A multilingual neural coaching model with enhanced long-term dialogue structure

ASIER LÓPEZ ZORRILLA and M. INÉS TORRES, University of the Basque Country UPV/EHU, Spain

In this work we develop a fully data driven conversational agent capable of carrying out motivational coaching sessions in Spanish, French, Norwegian and English. Unlike the majority of coaching, and in general, well-being related conversational agents that can be found in the literature, ours is not designed by hand-crafted rules. Instead, we directly model the coaching strategy of professionals with end users. To this end, we gather a set of virtual coaching sessions through a Wizard of Oz platform, and apply state of the art Natural Language Processing techniques. We employ a transfer learning approach, pretraining GPT2 neural language models and fine-tuning them on our corpus. However, since these only take as input a local dialogue history, a simple fine-tuning procedure is not capable of modeling the long-term dialogue strategies that appear in coaching sessions. To alleviate this issue, we first propose to learn dialogue phase and scenario embeddings in the fine-tuning stage. These indicate to the model at which part of the dialogue it is and which kind of coaching session it is carrying out. Second, we develop global deep learning system which controls the long-term structure of the dialogue. We also show that this global module can be used to visualize and interpret the decisions taken by the the conversational agent, and that the learnt representations are comparable to dialogue acts. Automatic and human evaluation show that our proposals serve to improve the baseline models. Finally, interaction experiments with coaching experts indicate that system is usable and gives raise to positive emotions in Spanish, French and English, while the results in Norwegian point out that there is still work to be done in fully data driven approaches with very low resource languages.

CCS Concepts: • **Computing methodologies → Discourse, dialogue and pragmatics**.

Additional Key Words and Phrases: Dialogue System, Coaching, Multilingual, Transfer Learning, Explainable Artificial Intelligence

1 INTRODUCTION

The application of dialogue systems or chatbots in healthcare and well-being is a rapidly growing research area. These conversational agents aim at improving some or many aspects of the users' health. For instance, they may be used to help treating diseases like asthma [45] or cancer [7], monitoring health related parameters [84], prevention and treatment of mental health disorders [14, 42], or to provoke reflection [47] and motivate healthy behaviour changes [75] to, e.g., increase the amount of fruit [11], control weight [41] or cease smoking [23]. These tasks differ considerably from the classical application domains of dialogues systems[71, 95], which have often been devoted

Authors' address: Asier López Zorrilla, asier.lopezz@ehu.eus; M. Inés Torres, manes.torres@ehu.eus, University of the Basque Country UPV/EHU, Leioa, Spain.

just to provide some information or service to the user, such as checking the weather or restaurant booking, or just chit-chatting. From the perspective of the dialogue strategy, there is a big difference between providing information or simple services and trying to, for instance, provoking behavioural changes. In the latter there is no rush to complete any task; it is more important to calmly converse with the user and make them aware of their own problems, obstacles and potential goals they may want to achieve. The objective of the health-related conversational agents being so different and delicate, the employed methodologies are also significantly different. Some works propose simple user interfaces such as, multiple choice inputs for their system or just question-answering systems. Even among those which allow a dialogue via (spoken or text based) natural language interface, the dialogue strategy is almost always implemented by a hand-crafted strategy or at least non fully data driven approaches, such as finite state or frame based management [50]. On the other hand, some of the most promising chatbots in open-domain dialogue modeling are solely based on machine learning and are fully data driven [2, 86], which is radically different to the aforementioned dialogue management approaches. In this work, we aim at bridging this gap, applying state-of-the-art Artificial Intelligence techniques to develop a conversational agent capable of carrying out coaching sessions. We propose several improvements to adapt open domain dialogue modeling techniques to the needs of behavioural change models and being able to effectively apply a dialogue strategy from the perspective of planned behaviour [3].

Unlike rule-based conversational agents, which are often implemented taking into account the consideration of experts, we will try to learn and model directly their professional coaching strategy. To this end we will use the data acquired within the project in Spanish, French and Norwegian. The dialogues were then translated into English, creating a multilingual corpus in four languages. The fact that the corpus is multilingual already poses a major challenge. The deep learning system we will propose in this work will be only word-based, i.e., we will try to model professional coaches without using any type of symbolic turn representation like dialogue acts or name entities. While this ensures that our approach can be easily replicated in other contexts and that does not require expensive labeling, it hinders our task, especially when working with very low resource languages like Norwegian. The second challenge to overcome is to build a conversational agent based on this technology that is capable of modeling complex conversations with long-term dialogue strategy like coaching sessions. This is especially difficult because deep learning approaches similar to the one proposed in this work have been mainly employed in very short dialogue tasks [104], or in open domain dialogue modeling where the long-term structure of the dialogue has been completely ignored [2], even though even social dialogues have an underlying structure [32].

To address these challenges we build upon a transfer learning approach, which has recently been adopted and proved to be successful in many dialogue modeling tasks [35, 104]. This methodology has turned out to be very handy and attractive in Natural Language Processing (NLP) in general, mostly due to big research teams releasing very large and pretrained neural language models such as GPT [79], GPT2 [80] or BERT [21]. These neural language models have shown to have a great generalization ability, and can be fine-tuned and converted into up and running generative conversational agents. In fact, experts in coaching have recently pointed out that it is necessary to research the applicability of these giant neural network models in well-being related tasks [107]. However, these models are mainly developed for the English language. Thus, we propose to pretrain such neural language models on big open domain text corpus available in many languages such as OpensSubtitles or Wikipedia, and then fine-tune them on our smaller coaching corpus.

On the other hand, the main point to be taken into account is that the target dialogues are coaching sessions. These, in contrast to open domain conversations, have a long-term structure that cannot be ignored, and therefore needs to be learnt. The open domain dialogue systems that we will take as baselines often take a local dialogue history only as input, and therefore, are unable

to keep long-term coherence. Thus, we propose two substantial methodologies to further adapt the models to our task. Our first improvement comes in the fine-tuning stage of the generative model. We propose to learn embeddings that indicate the model at which dialogue phase it is and which kind of coaching session is carrying out, so the generated responses are more coherent. Second, we propose to build an additional deep learning system which will be used to take into account the whole history of the conversation, i.e. the dialogue history. We will name it the Whole Dialogue History system (WDH system). The two models, i.e. the fine-tuned neural language model and the WDH system, will cooperate to produce a response as suitable as possible in the coaching environment.

The fine-tuned neural language model will act as a generative model which produces a set of candidate responses given the partial dialogue history. Thus, we will also refer to it as the short-term generative model. Ideally, if the training process has been successful, these candidates will be coherent short-term. They will take into account the current topic of the dialogue and the last information the user has provided. However, it may well happen that not all of the candidates are coherent long-term too. For example, the user and the agent might be talking about the user's dinner routine. Only taking into account that context, it might be reasonable to ask the user whether they take fruit at dinner time. However, the agent and the user might already have discussed about the fruit intake earlier in the dialogue in a way that it makes no sense to select this candidate as the final response. This is where the WDH system comes into play. It analyses (the contextual sentence embedding corresponding to) each turn in the whole dialogue history and computes a score measuring how suitable each generated candidate is. Following our example, this system would see that the agent and the user have already been discussing about fruit, so it would assign a very low score to that candidate, whereas other, more relevant and coherent candidates would be ranked much higher. Moreover, not only does the WDH system avoid repetitions, but it should also select, in general, candidates that follow more precisely the coaching dialogue strategy appearing in the corpus.

Additionally, we will also show that the WDH system can be a powerful tool understand and explain on what basis the decisions of the dialogue system are taken, which is an emerging concern in neural network based systems. In fact, we show that the unsupervised representations learnt by the WDH system are closely related to conventional dialogue acts, but with advantage that no costly annotations are needed to develop them.

Finally, we measure the impact of each of our proposals in terms of automatic metrics and human evaluation of the generated responses. We also provide an analysis of interaction experiments with our system in the four languages.

Thus, in summary, these are the contributions of this work:

- We develop a novel coaching conversational agent by directly modeling professionals. Our proposal is trained purely on text, no dialogue acts are used, which makes it more general and applicable in other domains. Additionally, it is multilingual, i.e., it is capable of carrying out coaching sessions in English, Spanish, French and Norwegian.
- We describe a novel approach to improve the quality and relevance of the candidates the fine-tuned neural language model generates. On the one hand, we use scenario embeddings to specify which scenario the model should carry out. On the other hand, we explain how to build dialogue phase embeddings, a simple and powerful resource to enhance a more fluid dialogue flow.
- We propose and validate a novel mechanism, the Whole Dialogue History system, to take into account the whole dialogue context to ensure the coaching model is coherent long-term.

- We also show that this system can be a valuable tool in terms of explainable Artificial Intelligence; it allows to visually analyze on what basis the system takes its decisions. To this end, we compare the learnt representations with dialogue acts too.
- Finally, we discuss the potential impact and acceptance the described system would have on real users, based on automatic and human evaluation of the system.

The rest of the article is organized as follows. Section 2 presents the related work. Section 3 provides more information about the coaching dialogues to be modeled and the acquired corpus. Section 4 gives a top level overview of the proposed system. Section 5 describes the short-term generative model. There we present our proposals for the fine-tuning stage, i.e. how to train scenario and dialogue phase embeddings. Section 6 describes the WDH system in depth. In Section 7 we give more details about the experimental setup; including information about the pretraining and fine-tuning of the generative model and training details of the WDH system's modules. We also describe the automatic metrics and human evaluation procedures. In Section 8 we report the results of these evaluations. Finally, in Section 9 we present the visual analysis to better understand the decisions taken by the system and present a comparison with dialogue acts. We conclude with a discussion of our findings and their implications in Section 10 and with some final remarks in Section 11.

2 RELATED WORK

2.1 Coaching conversational agents

Many diverse machine-assisted coaching systems, conversational agents and apps have been proposed in the last few years, forming a wide spectrum in terms of the employed technologies, implemented coaching methodologies and their area of application. In fact, besides healthcare and well-being, coaching systems with artificial intelligence (coaching AIs, in short) have recently targeted other domains such as leadership (e.g. PocketConfidant[1]) or employee training [66]. On the other hand, the coaching strategy also varies greatly. In this regard, it is important to mention that not all the coaching AIs in the market or in the literature make use of an NLP interface, and even less incorporate a conversational agent. Some, like HabitBull[2] or Remente[3], just track the user progress in one or many habits, and provide them with data analysis, motivational videos or interactive guides to motivate them to reach their goals. Others, such as Quenza[4] or Coach.me[5], also act as mediators between users and professional human coaches, allowing face-to-face online coaching sessions. However, since our work involves the design of a conversational agent, we are most interested in coaching AIs that approach coaching as a conversation between the coach and the coachee, or that at least contain a dialogue module inside them. We will first discuss some of the coaching chatbots that can be found in the market and then the works in the literature.

2.1.1 Coaching chatbots in the market. The so-called leadership bots, which aim at strengthening leadership skills, improving communication and developing self-confidence, have recently gained interest by many companies. Among this kind of coaching AIs, we can find PocketConfident[6],

[1]https://pocketconfidant.com/
[2]http://www.habitbull.com/
[3]https://www.remente.com/
[4]https://quenza.com/
[5]https://www.coach.me/
[6]https://pocketconfidant.com/about/

ROCKY[7] or LEADx Coach Amanda[8]. According to their websites, PocketConfidant *engages individuals in personal, private and meaningful conversations to get unstuck, develop and reinforce human competence*; ROCKY provokes reflection routines asking *questions to help you reflect or prepare on your day*, which *vary every morning and evening and get more personalized over time thanks to machine learning behind*; and the LEADx Coach Amanda is able to provide leadership tips and answers to employee problems. It seems to perform some kind of user customization too: *because the Coach Amanda HR chatbot knows your personality, she'll personalize your manager training down to the sentence level.*

Naturally, there are also coaching chatbots designed for health care and well-being related matters. Wysa[9] is one of the most notable chatbots in this regard. It has been awarded as the best health care app by ORCHA, and its effectiveness has been validated through clinical studies[42]. Wysa is able to keep relatively long dialogues with a mix of natural language and multiple-choice input, and uses cognitive-behavioral techniques to reduce the levels of depression and stress; fight frustration, loneliness, or isolation; and improve mental health in general. It has also been the first AI mental health app to meet clinical safety standards, more precisely, the NHS UK's DCB 0129 Standard of Clinical Safety. Youper[10] is another app for mental health that includes a conversational agent. It is designed to help the users overcome anxiety and depression, *applying behavioral coping skills, and monitoring mental health symptoms.* Youper has been listed among the top ten behavioral apps in terms of real-world stickiness and engagement [15]. Last, it is interesting to mention Replika[11], which acts more as a companion chatbot than an actual coaching AI. It is most popular among young people (its main users are aged between 18 and 25), and the authors claim that it can help managing the emotions, reducing anxiety and reducing sleeping troubles.

2.1.2 Coaching conversational agents in the literature. On the other hand, similar coaching AIs have also been proposed in the literature. As we will discuss, even if they include a conversational agent, most of them are more focused on the tracking and goal setting parts of the coaching rather than on the motivational conversations. Additionally, the described dialogue engines are not end-to-end. For example, [28] implement chatbot, CoachAI, which acts as a task scheduler and tracker to promote physical activity. To this end, it includes a dialogue engine that guides the user through a series of steps to achieve their daily goal. However, in contrast to the dialogue model presented in this work, theirs is not a data driven one. Instead, its core is a structured finite state machine. Another work that presents a coach AI to promote regular aerobic exercise is [70]. The users can set their weekly goal, and the system keeps track of it, schedules exercises and offers future goals depending on their progress. However, this coach AI does not use any complex conversational system to interact with the user: it relies on rule-based heuristics to drive the coach's reasoning. [31] present an ongoing work on building a conversational agent to perform conversations about daily living to determine their degree of independence and assessing them. While they mention their intention of building end-to-end dialogue models in the future, the described conversational agent relies on rule-based dialogue policies due to the lack of training data. [9] describe an interesting coaching system for insomnia therapy made of two modules, a conversational agent and a module in charge of data acquisition, analysis and visualization. The dialogue system, however, is rather simple and it is based on multiple choice inputs.

[7]https://www.rocky.ai/chatbot
[8]https://leadx.org/hr-ai-chatbot-coach/
[9]https://www.wysa.io
[10]https://www.youper.ai/
[11]https://replika.ai/

Closer to our domain of interest, nutrition, [16] describe a coaching chatbot to help people improve their food lifestyle. It offers two goal possibilities to the user: reduce their meat consumption or increase the amount of vegetables and fruit they take. Besides tracking the user's situation with respect to their goal, it also offers the possibility to have guided conversations with the coaching chatbot about some predefined topics. The agent is able to provide the user with relevant images, videos and links to illustrate its remarks. The dialogue manager is built with the Chatfuel[12] service, which allows to manually design dialogues using a graphical interface. Interestingly, this chatbot was deployed in French rather than English. [68] describe the results of a single-arm pre-post study carried out to test the efficacy of a virtual health coach focusing (Mediterranean) diet and exercise. They show that the use of the Paola chatbot was able to reduce the weight of the participants and highly increase their Mediterranean diet score [91]. IBM Watson Virtual Assistant artificial intelligence software was used to design and implement the dialogue system, which allows the chatbot to converse with a natural language interface. This module, in contrast to our approach, is based on intent and entity detection to provide an appropriate response from a set of predefined options.

Thus, our proposal is one of the very few works that describes a conversational agent capable of carrying relatively long dialogues with natural language input. Moreover, to the best of our knowledge, this is the first attempt to build a fully data driven end-to-end coaching conversational agent.

2.2 Multilingual or non-English end-to-end dialogue systems

There have been diverse attempts to build multilingual or non-English dialogue systems, yet the amount of works describing end-to-end[13] dialogue models based on neural networks is rather scarce. Due to the lack of conversational data in many languages, some authors tackle this problem using automatic translation systems to convert the input message into English, then use an English chatbot to generate a response and finally translate it back into the original language [81]. Nonetheless, there are also some few examples of end-to-end neural dialogue systems trained directly in other languages. For example, [17] presented a chatbot in Chinese and a multilingual version of it in Chinese and English based on memory networks. Generative Adversarial Networks have been used to train multilingual response selection systems [88] or response generation models in very low resource languages like Basque [62]. Closer to our transfer learning approach, [57] built a multilingual transformer capable of interacting in six languages other than English, trained on a multilingual version of the Persona-chat database.

Nonetheless, multilingual end-to-end dialogue systems is definitely a growing area of research. For example, many authors have recently targeted some cross lingual and dialogue related tasks, such as dialogue breakdown detection [56], intent detection and slot filling [10] or topic classification [72].

2.3 Control mechanisms to strengthen the long-term coherence of end-to-end dialogue systems

The task of keeping track of the dialogue context has been tackled since the early task oriented dialogue systems. When the objective of the dialogue is to fulfill a goal of the user, it is necessary to know how close to that goal the dialogue is. To this end, goal oriented dialogue systems have since then used a dialogue state tracking module. At first, a set of hand-crafted rules were normally used

[12] https://chatfuel.com/

[13] The term *end-to-end* is used with slightly different connotations by the machine learning community. In this work, with *end-to-end* we mean dialogue systems which produce a response based solely on the text corresponding to the dialogue history without using any kind of turn representations like dialogue acts or name entities.

to track the dialogue state. Afterwards, with the advent of POMDPs [103], probabilistic methods, such as dynamic Bayesian networks or attributed bi-automata [93], gained popularity also for dialogue state tracking. Since the revolution of deep learning, a variety of approaches to track the dialogue state and/or to take into account the whole dialogue history have been proposed in task oriented settings. Hybrid Code Networks [102], dialogue policies to specify actions plans [38], or, in general, pipelines that include a dialogue state tracking module have been proposed [33, 36, 59], among others. However, in all these cases the dialogue state and flow is controlled mainly or at least partially via dialogue acts extracted from the previous system and/or user turns. Therefore, all the methodologies require an annotated corpus (or hand-crafted rules) at some point to predict the dialogue acts. Our proposal does not.

Other works have tried to make use of the whole dialogue history in a similar manner to our approach, but with different goals. [6] used a recurrent neural network on top of turn embeddings to improve the dialogue act prediction. [96] integrate the whole spoken dialogue history using a variety of sentence embeddings for a semantic slot filling task. [76] present a dialog act classification system on automatically generated transcriptions that combines convolutional neural networks and conditional random fields for context modeling. [61] also perform a dialogue act classification via a hierarchical deep learning model that takes into account the dialogue context. [100] keep track of the dialogue history with a dual dynamic memory network and use it to make queries to a knowledge base in a task oriented setting. The work presented in [85] has been particularly inspiring for us. They propose to model the dynamics of turn embeddings to automatically evaluate the quality of the dialogue in the long run.

Nonetheless, to the best of our knowledge, this is the first work which models the dynamics of turn embeddings to include them in the decision making stage in a system which does not make use of dialogue acts.

2.4 Conditioning the output of generative networks

In respect of our proposal to learn scenario and dialogue phase embeddings, we can find related works in the literature that condition the output of generative networks in several ways, and with several purposes. Some examples include using Reinforcement Learning to control the repetitiveness of a model [87] or different approaches to assign the model a fixed personality [53, 108]. Conditioning the output of a generative model is very related the area of stylized response generation too. For instance, [30] propose a chatbot which generates responses in a similar style to a non conversational corpus. A similar approach has also been used to ensure a certain level of politeness in the system responses [74], and to generate system turns that express given emotional status [111]. However, none of this works use a similar strategy to the one presented in this work, to the best of our knowledge.

3 COACHING DIALOGUES AND CORPUS

3.1 GROW coaching dialogues

While there are several coaching techniques, in this work we will try to build a conversational agent that carries out dialogues leaded according to the GROW coaching model[14] [101]. The selection of this model was suggested by experts in the area, on the grounds that the GROW dialogues are much more structured than in other coaching methodologies[44]. This choice definitely permits an easier development of automatic dialogue systems which try to replicate the behaviour of the coach. A full GROW session consists on four main phases: Goals or objectives, Reality, Options,

[14]Mind that, in this article, we refer to coaching as behavioural change model which tries to make the coachees realize how they could improve their habits, not as an instructor or trainer which explicitly tell them what to do.

and Will or action plan. During the first phase, the agent aims at getting explicit objectives from the user, for instance, increase the daily fruit intake or reduce the amount of sugar in their diet. During the Reality phase, the coach and the user analyse the user's personal context and they try to detect potential obstacles that currently hinder achieving the previously established objectives. As for the Options phase, the goal is to generate and find options and tools that the user may use in order to face the aforementioned obstacles and achieve their objectives. In the last phase, the objective is to specify an action plan for the user to carry out in order to advance towards their objective. For instance, Table 1 shows a representative fragment from our corpus corresponding to the Reality and Option phases, where the coach provokes a clear reflection on the user. We will introduce the corpus in Section 3.2.

Table 1. A fragment of a coaching session extracted from our corpus.

Agent:	*So, Pablo, have you ever eaten consistently 2 or 3 pieces of fruit?*
User:	*When my wife was healthier, she used to take care of buying the fruit. Thus it was easier for me to eat it.*
Agent:	*And in any other time of your life?*
User:	*Long time ago, when I lived alone, I used to take care of buying the fruit myself, and I ate it more frequently.*
Agent:	*What does this information suggest to you about your objective?*
User:	*Well... That it's something that basically depends on me.*
Agent:	*So, can you see anything you could do to get closer to your objective?*
User:	*Uhmm... I should start thinking how I am going to organise to buy the fruit.*

Mind that this dialogue structure and strategy does fall into any of the two broad categories conversational agents are often classified in: task-oriented or transactional, and open-domain or social [19]. In the field of open-domain dialogue modeling there is no topic to talk about or task to carry out, the only goal is to generate appropriate and meaningful responses given a dialogue context. On the other hand, task oriented dialogue systems are often developed in order to provide the user with some information or service they request as soon as possible, such as hotel booking. On the contrary, GROW coaching dialogues have peculiarities that do not allow to easily classify them as transactional nor social. They are somehow task-oriented because there are some tasks to be completed, such as getting the user's objective or identifying which obstacles are not letting them fulfill their goal. However, the dialogue is definitely not carried out in a conventional task oriented manner. There is no rush to complete the task, and it is more important to calmly converse with the user and make them aware of their own problems, obstacles and potential goals they want to achieve. In this sense, coaching is also related to open domain dialogues, where there is no task and the only objective is to converse about different topics. However, coaching dialogues follow a clear and well structured strategy. These differences in the properties of the dialogues are the main reason why novel approaches and techniques are needed to model them.

3.2 Corpus

In order to model this dialogue strategy, a series of spoken dialogues have been acquired through a Wizard of Oz (WoZ) platform [90]. Two different scenarios were designed for the WoZ interactions. First, we designed an introductory scenario, which was used to engage the user and make them feel comfortable in the interaction with the system. In this scenario, the system presents itself and briefly describes the coaching methodology it will be following. Afterwards, it talks with the user about

their hobbies, such as travelling, music and reading, but always with the next coaching session in mind. Secondly, a (partial) GROW session on nutrition was simulated. Real GROW sessions often last for one hour or more, which we considered that could bore the users and make them feel uncomfortable. Instead, the session stopped around the 10 minute mark, which was generally enough to complete the first phase of the GROW structure, the Goals. Other times the conversation was very fluid, and more phases were completed. The nutrition topic was selected because it is a key factor for healthy ageing. According to the World Health Organization, "*good nutrition can help to preserve cognitive function, delay care dependency, and reverse frailty*" when ageing[15].

Dialogues were acquired in three different countries with different languages and cultures: Spain, France and Norway. A summary of the statistics of the corpus is showed in Table 2. Almost every user interacted with the system in the two scenarios, except some due to various reasons. Thus, the number of dialogues is slightly lower than the double of number of participants. Each dialogue was approximately 10 minutes long, which resulted in an average of roughly 29 turns per dialogue.

Table 2. A summary of the big numbers in the corpus.

	Spanish	French	Norwegian	Total
Number of participants	79	35	35	149
Number of dialogues	142	68	62	272
Number of system turns	4813	1776	1324	7913
System turns per dialogue	33.9	26.1	21.4	29.1
Running words	92K	47K	36K	175K

In order to increase the total amount of data in each language, all the dialogues were translated into the other two languages. This translation was done in a semi-automatic way: an automatic machine translator was used first, followed by a manual correction. English was used as an intermediate language for all the translations, due to the translators mostly being fluent in their language and in English. As a result, the corpus is also fully translated into English. Thus, the actual numbers for each language are the same and are the ones shown in the *Total* column in Table 2. While the acquired dialogues are spoken, we will only be using the textual transcriptions in this study.

This corpus has been gathered within the European Horizon 2020 Project EMPATHIC [64, 97]. In order to develop other modules not related to this work, all the turns in the dialogues were annotated in terms of topic, intent, name entities and emotions [43, 71]. Nonetheless, we will be using none of these annotations in this study, since we are most interested on working with unlabeled data and developing end-to-end neural dialogue systems. Hence, our research will also be potentially more general and helpful to others too, because not always corpora are labeled neither the labeled ones use the same label taxonomy. The corpus will be made publicly available after mid-2021, soon after project is finished.

4 SYSTEM OVERVIEW

We propose a dialogue system which can effectively model the described long-term dialogue strategies while dealing only with unlabeled text. The proposed system is made of two important parts: a short-term generative model which creates some response candidates given a local dialogue history, and a global module which ranks the candidates according to their relevance given the whole dialogue history. We will name this module the Whole Dialogue History system (WDH

[15]https://www.who.int/news-room/fact-sheets/detail/ageing-and-health

system hereinafter). Before getting into the details of both parts of the system and proposed novelties, we will provide a top level view of the system's functioning.

The short-term generative model is a fine-tuned neural language model, a GPT2 transformer more precisely. It is trained in a transfer learning fashion to produce responses similar to what a coach would, given the local dialogue history made of the last turns of the conversation. The responses are generated via a top-K sampling decoding, which allows the generation of many different candidates given the same local context. However, since the local history that the model sees is not large enough to take into account the coaching strategy we aim at modeling, some of the generated responses are likely to be non-relevant or inappropriate. In Section 5, we will propose some control mechanisms that can be included in the fine-tuning stage to alleviate this problem.

Nonetheless, we firmly believe that, in any case, it is necessary to take into account the whole dialogue history in order to successfully carry out complex dialogues like coaching sessions. If the model responses are produced only given the local context, repetitions might occur and in general non-consistent turns can appear very easily. Since with the current hardware it is not possible to include all the dialogue in the generative model's input due to memory limitations, we propose to build another system, the aforementioned WDH system. This evaluates how coherent each of the candidates proposed by the short-term generative model is given the whole dialogue history. The main idea behind the WDH system is to model the long-term dynamics of the dialogue and include them in the decision making stage. More precisely, we will model the path the dialogues follow in the abstract semantic space of sentence or turn embeddings. To this end, the embeddings will be grouped into clusters, with the assumption that turns inside each cluster should share some semantics and their role in the dialogue should not be too different. In fact, we will later show (in Section 9.2) that there is a strong correlation between the cluster a turn has been assigned to and the corresponding dialogue act. Figure 1 shows a bidimensional projection of the turn embeddings and the resulting clusters. For the sake of simplicity, the number of clusters shown in the image is lower than the actual one. For instance, it can be seen that the purple cluster in the figure contains introductory turns, such as greetings or system presentations; the black cluster turns about food routine; the green one is travelling related, and the light blue contains turns about music.

As it can be seen, the turn embedding space seems to be organized enough so as to provide valuable information in the decision making stage. We will discuss this space more in depth in Section 9. Note that, if we group each turn into a cluster, the dialogues in the corpus can be represented as sequences of clusters. Since the dialogues in the corpus follow certain patterns and strategies, these sequences should follow them too. We will try to model the sequences of clusters and produce a system response which belongs to a cluster that is likely given a certain cluster sequence. A diagram of the whole system is shown in Figure 2. In the diagram the GPT2 score represents the score that the generative model assigns to each candidate via a reranking procedure (more about this score on Section 5.4).

In addition to its relevance in the decision making stage, the WDH system can be employed to analyse and visualize the dialogues in the corpus. It also helps to explain and understand system's decisions. We show it at the end this study, in Section 9.

5 ADDING EMBEDDINGS TO THE SHORT-TERM GENERATIVE MODEL

In this Section we will focus on the short-term generative model that can be seen in Figure 2. The neural network trained to produce candidate responses given a local dialogue history is a sequence-to-sequence transformer model [98]. We will start with a pretrained GPT2 language model [80], and convert it into a response generation model applying transfer learning. In order to apply this methodology in the most effective fashion, it is key to exploit all the capabilities of the pretrained model. In the case of the GPT2 transformer models, [104] have already proved

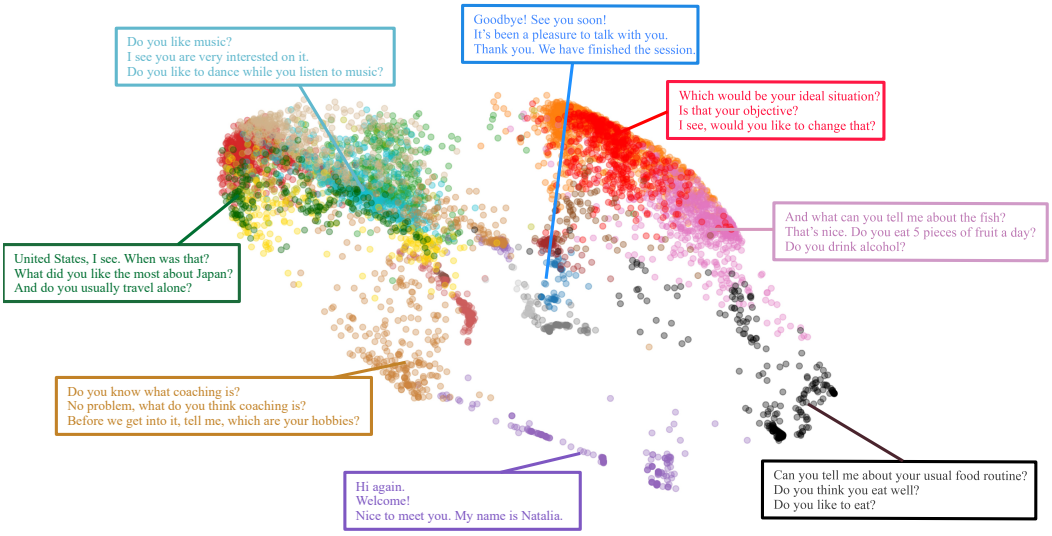


Fig. 1. A bidimensional projection of turn embeddings, coloured by the cluster they have been assigned to.

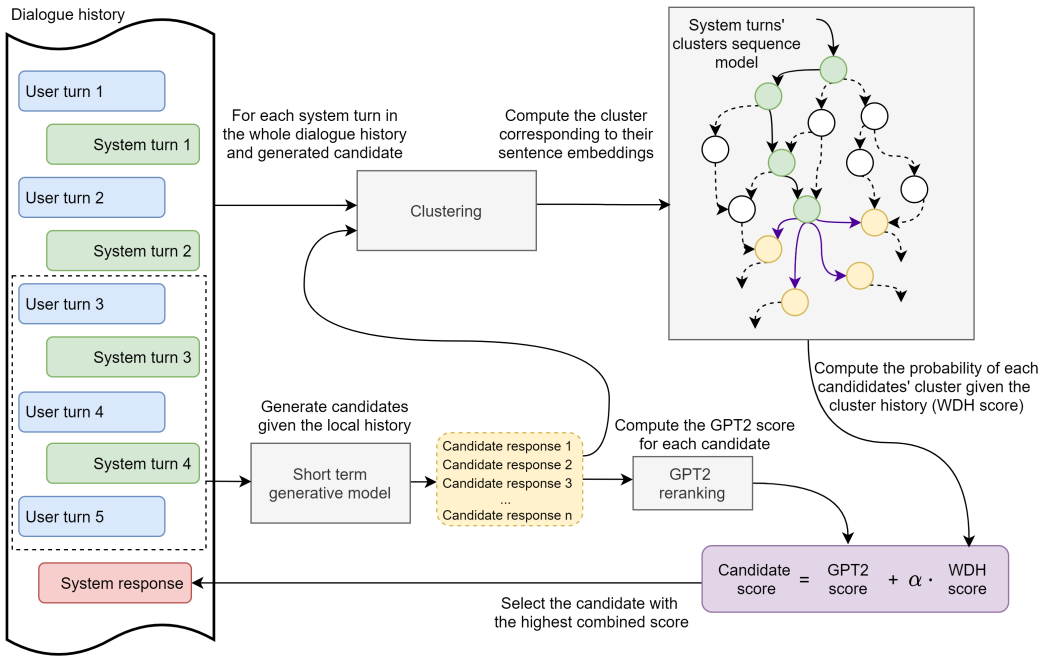


Fig. 2. The diagram of the proposed conversational agent.

that adding information in form of additional embeddings to the input representation can be very useful.

Thus, taking their work as baseline, we will introduce two modifications to the input representation to improve its performance and adapt it to the needs of a motivational conversation model. As

we will explain in Sections 5.1, 5.2 and 5.3, this proposal consists in learning different embeddings to control the behaviour of the network in one way or another. A diagram of an example of a complete input to our transformer can be found in Figure 3, where only two input turns are shown for simplicity. In Section 5.1 we explain our baseline model (BS), in Section 5.2 the scenario embeddings (SC), and in Section 5.3 the dialogue phase embeddings (PH). Finally, in Section 5.4, we give further details about the generative model.

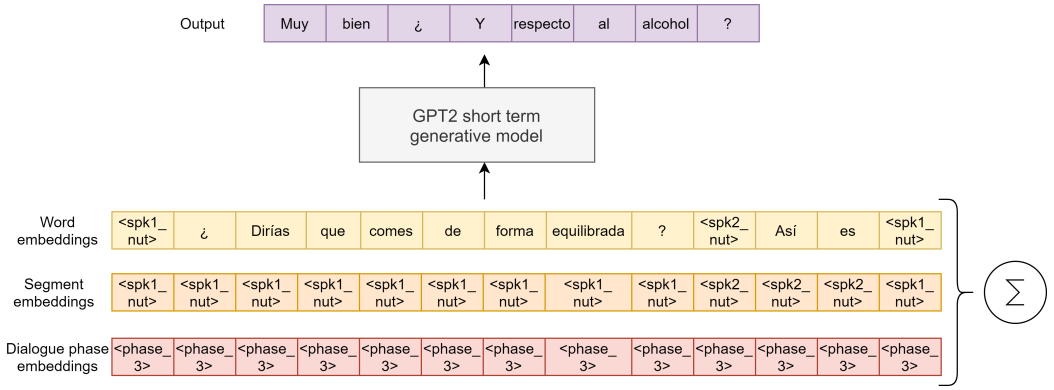


Fig. 3. An example of the proposed input representation to fine-tune the GPT2 transformer network. The actual input to the transformer is the sum of all the embeddings in each time step. The segment embeddings (Section 5.2) indicate that the system is performing a nutrition dialogue, and the dialogue phase embeddings (Section 5.3) that it is the third phase of the dialogue.

5.1 Baseline (BL)

In our baseline, the input is represented with two parallel sequences of embeddings[16]. Since the first layer of the transformer takes only one sequence of vectors as input, the embeddings corresponding to each time step are added before being fed into the transformer. Let us describe the task of each embedding.

The first sequence corresponds to the word embeddings of the word/tokens of last turns of the dialogues. In our experiments we use the last five turns in the dialogue history, and concatenate them using special tokens as separators. This sequence of embeddings is the first row in the example shown in Figure 3. The second one is used to segment the input into some categories. In our baseline, these segment embeddings indicate which input tokens correspond to the system's turns and which ones to the user's: <spk1> and <spk2>. This is the most straightforward way of applying transfer learning to convert a language model into a chatbot. While it might be interesting and appropriate for small conversations or just chit-chatting, in our case we need to ensure the overall robustness and coherence of the model if we want it to handle coaching sessions.

5.2 Scenario segment embeddings (SC)

First of all, we have to take into account that our task requires the dialogue model to be able to carry out two different kind of dialogues: an introductory dialogue and a partial GROW session about nutrition. Thus, we certainly need the option to specify which scenario to carry out to the model. It is also necessary that it does not arbitrarily jump from a scenario to the other. While we

[16] Of course, positional encoding embeddings are also used throughout the whole work, but they are omitted here for the sake of simplicity, because they are common to almost all the transformer networks [98].

could train two different models for each scenario to avoid these issues, this approach would not allow each model to benefit from the conversational patterns appearing in the other half of the corpus. We consider that training a single model with the whole corpus in a multitask fashion is highly advantageous in this situation where the amount of data is not very high.

We propose to substitute the segment embeddings of the baseline with four different segment embeddings, in order to indicate which type of dialogue to carry out to the model: <spk1_int>, <spk1_nut>, <spk2_int> and <spk2_nut>. These now indicate which the user is but also the scenario. For instance, the second row of Figure 3 the embeddings indicate that the selected scenario is the nutrition one.

The scenario segment embeddings provide consequently a way of controlling the topic the system will talk about with the user: if at the beginning of a dialogue we feed the nutrition segment embeddings, the model will then talk about nutrition. If, conversely, we use the introductory segment embeddings, the machine will carry out an introductory conversation. Furthermore, note that this idea can be easily implemented in many other multitask frameworks other than ours.

5.3 Dialogue phase embeddings (PH)

Finally, motivated by the empirical fact that the addition of (high-dimensional) embeddings is an appropriate technique to mix several pieces of information [104], we decided to add a third set of embeddings: the dialogue phase embeddings. This is devoted to enhance a dynamic progress of the conversation (without repetitions or loops) and a controlled ending. The phase embeddings tell the system at which point of the conversation it is, i.e., which proportion of the dialogue has been completed. For dialogues with lengths between 20 and 30 turns, we found that learning four dialogue phase embeddings was enough to lead to big improvements in terms of controlling the flow and limiting the length of the dialogue. Once a phase embedding is selected in function of the turn number and the desired length of the dialogue, it is added to all the input embeddings, as Figure 3 shows. Let us describe when each of the embeddings is used and which is its task, intuitively:

- The <phase_1> embedding is used in the first 20% of the dialogue. It tells the system that the conversation is starting, and thus when this embedding is added to each of the word embeddings, the systems tends to produce opening sentences or greetings.
- The <phase_2> embedding is used from the 20% of the dialogue until the 50%. It corresponds to the rest of the first half of the dialogue, where the system tries to find an appropriate topic of conversation, asking the user some open questions.
- The <phase_3> embedding is used from the 50% of the dialogue until the 90%. In this phase the system and the user mostly discuss about the topic they started in the second phase.
- Finally, the <phase_4> embedding is used within the last 10% of the dialogue. The system ends the discussion held in the previous phase, closes the conversation and says goodbye to the user.

We also investigated and tested other smoother designs for these embeddings, such as using different embeddings per each turn. Nonetheless, we ended up discarding this option because our corpus (as many others) includes dialogues of very diverse lengths: sometimes the conversation ends in turn 15 whereas other times at turn 15 the user is still starting to talk about their nutrition habits. This can definitively lead to these embeddings not being trained precisely, and hence we opted for the relative phase embeddings approach.

This one, besides being more suitable in this case, also introduces the option of manually selecting the desired length of a conversation once the model is trained. This control, albeit not being extraordinarily versatile, is enough to tweak the flow of the dialogue, which is very useful

when dealing with end-to-end neural dialogue models, where controlling the system responses is often a very tough task.

5.4 Decoding in the short-term model and GPT2 candidate reranking

In order to complete the description of the generative network of our system, let us now give details about how the decoding, i.e. the generation of candidates is carried out. We will also mention how the GPT2 score for each generated candidate is computed. In Figure 2, while the described input embedding proposals have been centered in the input arrow to the short-term generative model block, the decoding refers to its output arrow, and the GPT2 score is shown in the bottom right purple block, where the total score for each candidate is computed.

Decoding details: Neural dialogue systems have been well known to generate too generic and repetitive. This problem has been tackled with many approaches, such as modifying the loss function [52] or using adversarial training [54, 63]. Lately, making use of a proper decoding procedure has proved to be essential for generative models to produce good quality non-generic responses [34, 49]. We adopt the recently introduced nucleus sampling strategy [40] to prevent the system from generating dull or generic responses as much as possible. This technique consists in sampling only from a subset of tokens at each generation step. This subset is composed of the tokens whose cumulative probability is greater than or equal to a threshold. We set this threshold to 0.9. Additionally, prior to computing the aforementioned subset of candidate tokens at each generation step, we also apply some temperature [1, 29] to the logits to control the diversity of the responses. For our application, we found that temperatures ranging from 0.65 to 0.8 led to very interesting responses. The value we set for the final experiments is 0.7.

Candidate reranking via the GPT2 score: GPT2 models are often trained both to generate candidates given a context and also to predict the next utterance given a set of possible ones [36, 104]. More precisely, they are trained to predict whether a certain candidate is the correct response given the context or not. This binary prediction is done by a linear classifier that takes as input the hidden state of the transformer after processing the last token of the candidate. The output of this linear layer, i.e., the unnormalized probability of a candidate being the correct next utterance, will be the GPT2 reranking score. Intuitively, this score should be high when a candidate is informative and coherent with the local context; whereas non-relevant candidates or candidates containing grammatical errors should be assigned a low GPT2 score.

6 RERANKING USING THE WHOLE DIALOGUE HISTORY

Let us now present the WDH system in depth. It is composed of four modules. The first one's function is to produce sentence embeddings of each system turn. The second one carries out a dimension reduction of the previously computed sentence embeddings. The third one is a clustering module which assigns a cluster to the lower-dimensional embedding. These first three modules correspond to the *clustering* block in Figure 2. Finally, the last module produces an (unnormalized) probability distribution over all the possible clusters given the sequence of clusters that represents the dialogue history. This probability is the WDH score. In Figure 2, this fourth module is the block shown in the top right, and the resulting WDH score can be found the bottom right.

6.1 Contextual turn embeddings

There are several techniques to produce sentence embeddings, and each of them have shown strengths and weaknesses depending on the NLP task they have been employed in. In preliminary experiments we compared generic sentence embedding methods, such as multilingual universal sentence encoders [106], sentence transformers [83], or a weighted average of word vectors [5].

However, none of these methods worked as well as using the embeddings produced by the short-term generative GPT2 models. The embeddings are the hidden state of the transformer after processing the last token of the sentence.

Using the short-term model for computing these sentence embeddings not only simplifies the system's pipeline, but it also provides additional benefits. First of all, since the model takes as input a partial dialogue history, the embedding it outputs contains information about both the user and system turns. This should definitely be considered an asset, because it allows to pack the information of the user's turn in the system's contextual embedding. Otherwise, if a non-contextual embedding method were to be used, it would require to compute and process two different embeddings, one for the user and another one for the system. The second benefit of using a fine-tuned model is that the resulting embeddings are domain specific too, which is key for a better performance.

Since we definitely want the embeddings to include the scenario and dialogue phase scenario information mentioned in Section 5, we first considered using the full system with scenario and dialogue phase embeddings. However, further experiments showed that it was more convenient to use just the baseline model, and include the scenario and phase information in the dimensionality reduction stage we will describe next.

6.2 Dimensionality reduction

We apply a dimensionality reduction technique prior to the clustering method to avoid curse of dimensionality [8], since it is known to improve the quality of the clustering methods when these are distance or similarity-based [94].

We tried many methodologies such as PCA to carry out this dimensionality reduction, but we ended up building an autoencoder [48]. The reason for this is that, as aforementioned, we can easily incorporate supervision in the dimension reduction process, in a similar fashion to [51]. The most straightforward way of training autoencoders is to optimize a recovery loss from a space with a lower dimension than the original space. In addition to the recovery loss, we minimize two classification losses, computed after a linear transformation of the low dimensional space. These correspond to the scenario and dialogue phase classification.

A summary of the structure of the autoencoder can be found at Figure 4. It takes as input a sentence embedding x, and after applying some non-linear layers with successively less output size, it ends up transforming it into $\boldsymbol{h}$, the low-dimensional representation of $\boldsymbol{x}$. This is the vector we will be using at the clustering stage. Then additional layers transform $\boldsymbol{h}$ into $\boldsymbol{x'}$, the reconstructed version of $\boldsymbol{x}$. Thus, $\boldsymbol{h}$ contains as much information of $\boldsymbol{x}$ as possible. Furthermore, two linear layers perform two classifications from $\boldsymbol{h}$. After respective softmax normalizations, $\boldsymbol{y}_{scenario}$ and $\boldsymbol{y}_{phase}$ are produced, the probability distributions over the possible scenarios and dialogues phases. These two classifications ensure that the low-dimensional representation of the turn embeddings maintain as much information as possible about the scenario and dialogue phase, which are key properties of the turns.

Therefore, the training objective for this autoencoder will be a combination of three losses, as shown in Equation 1. We tried some weighted sums of the losses instead of the unweighted one, but we found no improvement. The reconstruction loss is the euclidean distance between $\boldsymbol{x'}$ and $\boldsymbol{x}$. On the other hand, $\mathcal{L}_{scenario}$ and $\mathcal{L}_{phase}$ are cross entropy losses for classification.

$$\mathcal{L}_{autoencoder} = \mathcal{L}_{reconstruction} + \mathcal{L}_{scenario} + \mathcal{L}_{phase} \tag{1}$$

6.3 Clustering the turn embeddings

After the sentence embeddings corresponding to the system turns are computed and dimensionally reduced, we propose to group them into clusters, in an unsupervised fashion. Intuitively, system

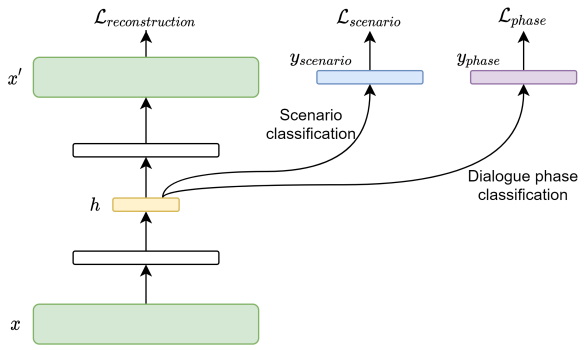


Fig. 4. A diagram of the proposed supervised autoencoder to reduce the dimension of turn embeddings.

turns that are close to each other in the low dimensional embedding space will be semantically close, and they will also share key dialogue information, such as the scenario and dialogue phase.

There are many techniques to perform unsupervised clustering, and which is the superior one if often a matter of the use case [89]. We tried and compared various methods, such as DBSCAN [27], Birch [110], OPTICS [4] and K-Means [67]. After an inspection of the turns inside each cluster, we decided to stick to the K-Means, because we found no improvement with the more sophisticated methods. Additionally, the K-Means algorithm provides two substantial benefits in our case. First, it takes the number of clusters as a parameter, which is very valuable for our application: we want enough clusters so that each of them represents a different state in the dialogue; but if the number is too large compared to the number of dialogues in our corpus, the task of learning the probability of the next cluster would not be feasible. A detailed analysis of the effect of the number of clusters is provided in Appendix B. The second benefit is that, in contrast to many other clustering algorithms, it allows to predict the cluster corresponding to a new sample in a very simple way. This is necessary when interacting with the system, because it is not possible to know the cluster a given turn corresponds to beforehand. Instead of having to train an additional classifier that learns to map from turn embeddings to clusters, the distance from the new sample to the cluster centroids can be measured, and the argument of the minimum will be the corresponding cluster.

6.4 Learning the next cluster probability distribution

We cast the task of learning the next cluster probability as a sequence modeling problem. Given a set or vocabulary of m clusters $\mathcal{V} = \{v_1, v_2, ..., v_m\}$, and a sequence of clusters corresponding to a dialogue history $c_1, c_2, ..., c_n$, the objective is to compute the discrete probability distribution of each cluster being the next one in the sequence (Equation 2).

$$P(c_{n+1} = v_i \mid c_1, c_2, ..., c_n), \ \forall v_i \in \mathcal{V} \tag{2}$$

This task is very similar to a language modelling task, but having clusters instead of words. Therefore, we considered classical language modelling methodologies to tackle this problem. Even though N-gram models are simple models and have broadly been used to this end, recurrent neural networks, GRUs [18] more precisely, were our final choice. The main problem with N-grams is that they are based on the Markov Assumption, which assumes that the probability of the next cluster can be computed based only on the last few clusters. We really want the WDH system to take into account the whole dialogue history, so the N gram models were finally discarded. On the contrary,

GRUs process the whole cluster sequence. Appendix A shows that, indeed, taking into account the whole sequence is highly beneficial, since the GRUs outperform the N-gram models in terms of accuracy and top N accuracy.

The objective function used to train the GRU was the negative log likelihood at the cluster level:

$$\mathcal{L}_{GRU} = \frac{1}{|C|} \sum_{c \in C} \frac{1}{|c|} \sum_{i=1}^{|c|} -\log P(c_{i+1} \mid c_1, c_2, ..., c_{i-1}) \tag{3}$$

where C is the training corpus made of sequences of clusters $\boldsymbol{c}$, each of them corresponding to a dialogue. The probability of c_{i+1} being the next cluster given the partial sequence $c_1, c_2, ..., c_{i-1}$ is computed by a softmax normalization on top of a linear classifier given the last hidden state of the GRU after processing the cluster sequence.

6.5 Computing the total score for each candidate

The GPT2 score and the WDH system's score are fused in a simple way. The total score is a weighted both scores, as shown in Equation 4. This Equation is also shown in the bottom right of Figure 2. A detailed analysis of the chosen value for the hyperparameter α and its role in the system's performance is shown in Section 8.1.

$$\text{Total score} = \text{GPT2 score} + \alpha \cdot \text{Cluster score} \tag{4}$$

7 TRAINING DETAILS AND EXPERIMENTAL SETUP

In this section we give more details about our implementation and introduce the experimental setup that was used to produce the results we will present and discuss in Section 8. According to our proposal, we will mainly be training and comparing six models:

- The baseline model (BL). This refers to the model presented in Section 5.1, without any reranking. I.e., here we will only be using the short-term generative model, without the WDH system. This will generate just one candidate and it will be used as the system response.
- The baseline model with scenario embeddings (BL+SC). In this model we add the scenario embeddings (Section 5.2) to the baseline model.
- The baseline model with dialogue phase embeddings (BL+PH). Here we add the dialogue phase (Section 5.3) to the baseline model, but without the scenario embeddings.
- The full generative model (FM). This one includes both the scenario and dialogue phase embeddings. But still there is no reranking, i.e., it outputs the first utterance it generates.
- The full model with just GPT2 reranking (FM+RR). In order to check the influence of the WDH system, we first include a reranking process with only the GPT2 score. We generate and rank 10 candidates.
- The full model with both the GPT2 reranking and the WDH reranking (FM+WDH). This model includes all our proposals. The number of candidates is the same than in the FM+RR model, 10.

We first explain the process of pretraining the GPT2 neural language models in Spanish, French and Norwegian. Then we get into the fine-tuning details of these models with the EMPATHIC corpus. Finally, give details about the WDH system and introduce the experiments and the evaluation procedures.

7.1 Pretraining procedure

We are dealing with a multilingual corpus in Spanish, French, Norwegian and English. However, most of the big pretrained neural language models are only available in English. After some

preliminar experiments using multilingual pretrained transformers such as XLM [20], we found that fine-tuning these did not result in great dialogue models. Thus, we ended up pretraining GPT2 models from scratch in Spanish, French and Norwegian, and using the pretrained and freely available GPT2 models in English.

There are four different GPT2 architectures [80], which mainly differ in the number of layers and their size. We selected the *small* GPT2 transformer architecture for all of our experiments in Spanish, French and Norwegian, which has 124 million parameters. This selection was made to meet two important criteria: the model should be large enough to be capable of learning our task, but small enough to fit into standard GPUs and be pretrained in a reasonable amount of time. As for English, we compared the *small* model with the *medium*, which has 324 million parameters. The latter one worked much better already since the first experiments, as shown in Section 8.1. Therefore, unless it is mentioned explicitly, the results for the English system were achieved with the *medium* GPT2.

We used the Spanish, French and Norwegian versions of Wikipedia and OpenSubtitles [58] to pretrain each language model. The reason for choosing these corpora is that both are available in the target languages, and that both include valuable information which could improve the final performance of the coaching dialogue model. Wikipedia contains information about millions of topics, and OpenSubtitles is made of conversations mainly, which hopefully helps the model learning dialogue skills. Since the amount of data in Norwegian was much lower, we also included a fraction of the Norwegian version of the OSCAR text corpus [77]. OSCAR is a subset of Common Crawl, and thus it is made of web scrapped text from the Internet. Mind that, therefore, this corpus may not be as related to our task as OpenSubtitles or Wikipedia. Table 3 shows a summary of the data used for pretraining after having cleaned lines containing strange characters, urls, and so on.

Table 3. Statistics of the corpora used to pretrain the GPT2 model in Spanish, French and Norwegian. In Norwegian, values in brackets refer to the data prior to the addition of a fraction of the OSCAR corpus.

	Spanish	French	Norwegian
Amount of raw text	10GB	7GB	5GB (1GB)
Number of sentences	230M	121M	30M (14M)
Running words	1.7B	1.3B	750M (150M)

We first trained a BPE tokenizer [92] in each language with this data. This tokenizers will be used during the pretraining and fine-tuning steps. We selected a vocabulary of 10K subwords in each language. This number is slower than the pretrained tokenizer in English, which has a vocabulary of around 50K subwords. Using a reduced vocabulary size reduces also the memory consumption and training time.

We then trained each GPT2 model from scratch, throughout two complete epochs on each dataset. We set the maximum number of input tokens to 512, which we consider enough since it allows us to afterwards include 5 turns of dialogue history in the fine-tuning step. We used the ADAM optimizer [46] with a linearly decaying learning rate from 1e-5 to 5e-4, and a batch size of 4, the maximum that fitted in our GPU. We clipped the gradients at a maximum absolute value of 1. Each training procedure took around 2-3 weeks in total to be completed in a single Nvidia Titan Xp GPU.

7.2 Fine-tuning the GPT2 generative model on the EMPATHIC corpus

After pretraining the language models, we fine-tuned them on our dialogue corpus to convert them into dialogue models. We fine-tuned each model with combinations of the three input

representations explained in Section 5, for comparison purposes. We trained the baseline, the baseline with scenario embeddings, the baseline with dialogue phase embeddings, and finally the full generative model with both scenario and dialogue phase embeddings. All the systems were also trained to predict the end of the dialogue. To this end, an end of dialogue token was inserted in the last system turn of every dialogue. The number of turns selected for the local dialogue history was five: three user turns and two system turns. In order to measure the effect of not pretraining, we also trained a GPT2 model from scratch with our corpus in Spanish.

We split the data into train (85%) and test (15%) partitions. These proportions were kept when splitting the original dialogues in each language and also the dialogues translated from the remaining two languages. Each partition also contains the same number of introductory and nutrition dialogues. Since most of the users interacted with the system in both scenarios, we also made sure that all the dialogues corresponding to a given user only appeared in one of the partitions.

Training details: Following previous work we employ multitask learning to fine-tune our network; optimizing a linear combination of two loss functions during the fine-tuning step [13, 79, 104]: the language model loss and the next turn prediction loss. The second one also enables the possibility of using the GPT2 score described in Section 6.5. We set the weight of the language model loss to be the double of the next turn prediction one. We used 10 candidates for the next turn prediction loss, the actual ground truth, 3 system turns from the previous dialogue history (but not appearing in the local history), 3 system turns that occurred later in the dialogue, and 3 random turns sampled randomly from the training set. The combined loss function was minimized throughout 4 epochs via the AdamW optimizer [65]. The learning rate was linearly decreased from 6e-5 to zero, gradients were clipped at their absolute value of 1 and weight decay of 0.01 was used. We could only fit one training example at a time in the GPU during the training process, but we still experimented with greater virtual batch sizes, accumulating the gradients for some steps. We found that a virtual batch size of 4 led to the most consistent results.

Additional details: The desired length of the dialogues was fixed to 20 system turns. As for the systems that use reranking (BL+RR and BL+WDH), in the decoding step we generated and ranked 10 candidates.

7.3 WDH system details

As mentioned in Sections 6.1, 6.2 and 6.3, we considered many strategies and algorithms to compute the sentence embeddings, dimension reduction and clustering. The final choice for each module in the pipeline was decided after an inspection of the resulting clusters. We checked that the turns grouped in the same clusters were in fact semantically close, and that it would make sense to use them in similar dialogue contexts. Finally, the baseline short-term generative model was used to produce sentence embeddings, a supervised autoencoder for dimension reduction and the K-Means algorithm for clustering. In Section 9, we provide a more insightful analysis of the results of the clustering pipeline.

Let us now give the details about the architecture and hyperparameter selection in the WDH modules. The turn embeddings were computed with the BL model. As for the autoencoder, its input and output size is the same than the turn embeddings. In the case of the Spanish, French and Norwegian systems this was 768, and in the case of English 1024, due to the use of the *medium* GPT2 architecture instead of the *small* one. The autoencoder's encoder and decoder are symmetrical. They are made of three fully connected layers of sizes 200, 50 and 5. The hyperbolic tangent was used as the activation function. Thus the low dimensional embedding space is of dimension 5. The two classification layers take as input this embedding and linearly classify the scenario and dialogue phase. The autoencoder was trained with the sentence embeddings of the training set of

the corpus during four epochs via the Adam optimizer. A batch size of 4 and a learning rate of 1e-4 were used.

As for the clustering, the Elkan's variation of the K-Means algorithm was used [24], with the euclidean distance in the low dimensional embedding space. After analysing its impact on different metrics, the number of clusters was set to 60. As shown in Appendix B, this value represents a nice compromise between a balanced number of turns per cluster and the performance of the WDH system at the next utterance classification task (which we will introduce next in Section 7.4). Additionally, it is also a value that permits a good mapping from cluster index to dialogue act, as explained in Section 9.2, where the correlation between the clustering and dialogue act classification is explored.

Once the clustering pipeline was fixed and trained, we proceeded with the cluster sequence modeling experiments via GRUs. The hyperparameters of the recurrent neural network were tuned in a development partition within the training set to preserve the train-test independence. The input size of the cluster sequence modeling GRU was set to 5. Namely, each cluster was represented by a five dimensional vector. We tried initializing them in terms of the turn embeddings but got no improvement, so they were randomly initialized and learnt in the process. Two GRU layers of hidden size 60 were used following, and finally a softmax layer of size 60 was used to output the probability distribution over the possible 60 clusters. The GRU was trained during 3 epochs via the Adam optimizer, with a batch size of 4 and a learning rate of 1e-4. The results in terms of accuracy and top N accuracy at the next cluster prediction task for each model and language are shown in Appendix A.

7.4 Automatic and human evaluation

Once all the models were trained on the train partition of the corpus in all the languages, we evaluated each of them according to their responses in the test partition. On the one hand, we computed some automatic metrics to measure the similarity of the generated response to the ground truth and the accuracy of the reranking methodologies. On the other hand, experts in coaching compared the responses of different models and selected the most appropriate one. Finally, these experts also interacted with the best model and evaluated the resulting dialogues.

Automatic metrics: Automatic evaluation of dialogue models is a very active and complex research area. In the last few years, many authors have been seeking metrics that measure the quality of the responses, and that correlate as much as possible with human evaluation in terms of, e.g. relevance, semantical appropriateness or informativeness. On the one hand, there are word overlap metrics such as BLEU [78], which measure how the generated response and the ground truth resemble at the word level. More recently, this similarity has also been measured via word or sentence embeddings [109]. There are even authors who propose unsupervised metrics which do not rely on ground truth responses [69, 73].

In this work we use two of the official metrics proposed in the The Conversational Intelligence Challenge 2 [22]: the accuracy at selecting the correct next utterance among a set of 10 candidates or next utterance classification accuracy, and the F1 score between the set of words of the response generated by the system and the ground truth. Additionally, we include the BLEU score as an additional measure of how similar the ground truth and the generated response are.

The next utterance selection accuracy is particularly interesting in our case, since much of our work focuses on improving the selection of good responses given a set of candidates. Note that this metric does not directly evaluate the response generation task. Instead, it focuses on the ability of the different models on selecting the correct response from a set candidates sampled from the corpus. This selection is done via the aforementioned GPT2 reranking modules (Section 5.4), and also with the WDH system in the case of the FM+WDH model. In any case, this metric should be a

nice indicator of the systems' performance when interacting with real users. The only difference is that in that case the set of candidates are not sampled from the corpus, but generated by the generative model. In the original metric of the Conversational Intelligence Challenge 2 [22], the set of candidates is made of randomly sampled responses entirely. However, in our case, 6 out of the 10 candidates are system turns that are part of the same dialogue, which makes the task more challenging since many candidates will probably be closer semantically. Among the remaining candidates, 3 are randomly sampled from the corpus, and last one is the ground truth.

Human evaluation: On the other hand, we carried out two series of human evaluation processes to validate our methodologies in the task of coaching. Since coaching is not a simple topic and expertise is needed to evaluate how good a system would be in this area, the selected human evaluators were the same professionals that carried out or participated in the Wizard of Oz experiments to acquire the corpus. This is very important, because it may well happen that a non-expert human considered that the interaction with the system has been good for example, but that would not ensure that the system is actually performing any type of coaching.

In the first series of evaluations we evaluate response quality of the different versions of our model. In order to measure the impact of each proposal, we compare them in an incremental fashion, through a sequence of pairwise comparisons. We start comparing the BL with the BL+SC to check the influence of the scenario embeddings. We do the same with the dialogue phase embeddings, comparing the BL with the BL+PH. Then the influence of adding the both embeddings is measured via the comparison of the BL with the FM. Afterwards, we analyse the impact of the candidate reranking with two comparisons, FM vs. FM+RR and FM+RR vs. FM+WDH. Finally, to give a grasp of the absolute quality of the responses, we compare the FM+WDH with the ground truth (GT hereinafter), i.e., the the human response found in the test set. Note that reason behind the choice of carrying out these incremental comparisons pairwise instead of, e.g. via a likert-score based evaluation per model, is that the results would be harder to compare, due to the potential evaluator bias when selecting the score in the likert scale, as discussed in [55]. Some annotators might be more generous and while others might tend to stick to more neutral responses. That bias is reduced in a pairwise setup, because the evaluators should only select which answer is better (or whether they are equal), but not to what extent. The biggest drawback of the pairwise comparisons is that might be difficult to aggregate the results if the comparisons are not incremental. In our case, this only happens with the BL vs. BL+SC and BL vs. BL+PH. This is why we also perform the BL vs. FM comparison.

In the second series of human evaluations, we focus on the usability and potential impact of the best system. We asked each coaching expert to interact with the model in each scenario and then to fill two questionnaires. Even though the system is planned to be used with a spoken interface, it was tested on a text based interface to avoid potential biases created by third modules. The first questionnaire is the chatbot usability questionnaire (CUQ) [39]. This novel questionnaire is similar to the classical system usability scale (SUS) [12] for human-computer interfaces, but adapted to the particular domain of chatbots, taking into account their peculiarities. On the other hand, the second questionnaire is based on the standarized questionnaire AttrakDiff [37]. AttrakDiff was designed to measure the user experience in human-machine interaction in four axis: the pragmatic attractiveness and three hedonic qualities. [26] adapted this questionnaire for the evaluation of virtual agents. In this study we will use the questionnaire related to one of the hedonic qualities axis, to the hedonic quality stimulation or feelings, more precisely. It aims at identifying the feelings may arise on the user when interacting with the system. This is particular important to assess the usability and potential consequences of a health-care related conversational agent. A system which gives raise to negative feelings on the user would never be acceptable, for instance. We will refer to this questionnaire as HFQ (Hedonic Feelings Questionnaire). Both questionnaires can be found in

Tables 14 and 15 in Appendix C. The responses were arranged in a five level Likert scale ranging from *Strongly agree* to *Strongly disagree*. Since both questionnaires ask about positive qualities of the system in even questions and about negative in odd ones, a score for each questionnaire can be easily calculated. A score of 100 would be obtained if a evaluator would *Strongly agree* with all the positive questions and *Strongly disagree* with all the negative ones, and a 0 in the opposite case.

8 RESULTS

In this section we present and discuss the automatic and human evaluations of our proposal.

8.1 Automatic evaluation

Let us now show the results of the automatic metrics. We will start discussing the performance of the models in terms of the next utterance selection accuracy among 10 candidates (Table 4), since it provides the most consistent results across all languages. We will then provide the results in terms of F1 and BLEU scores.

Table 4. Next utterance classification accuracy among a set of 10 candidates obtained by all the models in the four languages in the test partition of the corpus.

	Next utterance classification accuracy			
	English	Spanish	French	Norwegian
BL no pretraining	-	0.251	-	-
BL small	0.461	-	-	-
BL	0.482	0.374	0.404	0.350
BL+SC	0.488	0.379	0.402	0.343
BL+PH	0.488	0.388	0.417	0.366
FM	0.494	0.401	0.421	0.375
FM+RR	0.494	0.401	0.421	0.375
FM+WDH	**0.518**	**0.412**	**0.435**	**0.388**

Next utterance classification accuracy: First of all, we can see that there is a big gap between not pretraining the baseline and pretraining it in the Spanish model. Given this big gap, we did not consider trying with non pretrained baselines in the other languages. In English, there is also an improvement if we consider the medium GPT2 architecture (BL) or the small one (BL small). Therefore, the rest of experiments were carried out with the medium architecture.

Including the scenario embeddings does not seem to influence this accuracy as much as including the dialogue phase embeddings do. This is probably due to the nature of the candidates to be ranked: among the 10 candidates, 6 are system turns of the same dialogue. This can probably confuse the BS and the BL+SC more than the BL+PH model, because the candidates share the scenario and potentially the topic, but the phase embeddings might be able to capture that they are out of position given the status of the dialogue. The FM further improves over both BL+SC and BS+PH, proving that combining both embeddings leads to a better performance. In respect of the FM+RR, note that the next utterance accuracy is the same than the FM model. This is because these models are essentially the same; they only differ in the decoding stage: the FM generates just one response, whereas the FM+RR generates a number of candidates and then selects the best according to the GPT2 score. But in this case, since the set of candidates are given, there is no big difference. Finally, the full model with the WDH reranking method obtains the best results across all the languages. This clearly shows that the proposed reranking method helps to improve the candidate selection

.

criteria. Consequently, it also reinforces our initial hypothesis that it is necessary to process the whole dialogue history to improve the overall quality of end-to-end neural dialogue systems. This is even more critical when no dialogue acts or dialogue state tracker are being used; and also when the application, such as coaching, requires the dialogues to be well structured.

Let us show further the influence of the WDH system in the next utterance classification accuracy. As explained in Section 6.5 and Equation 4, the total score for a response candidate is a weighted sum of the GPT2 score and the WDH score, where α is the weight of WDH score. We performed a grid search with values of different orders of magnitude for this weight. The results are shown in Figure 5.

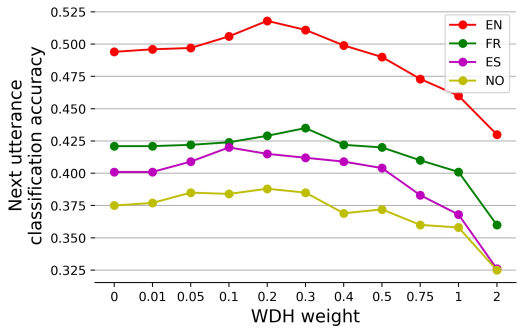


Fig. 5. The next utterance selection accuracy with respect to the WDH score's weight. Note that the x-scale is equally spaced between the tested values.

In general, the behaviour of the metric as a function of the cluster score is the expected one. When the next cluster score is 0, the model is equivalent to the FM+RR, and so the next utterance selection accuracy are the same shown in Table 4 for the FM+RR models. On the other hand, the accuracy peaks when the next cluster score is between 0.1 and 0.3. The maximum values are the ones shown in the table. If we increase the weight of the cluster score way much than its optimal value, the accuracy decreases drastically. This means that the WDH system should be used only as an addition to the GPT2 score, not as a substitute. The reason for this is that the GPT2 score takes into account properties that the WDH does not, and vice versa. The GPT2 score focuses on short-term coherence, but also in grammatical appropriateness, since it evaluates the turn embedding. On the other hand, the WDH is not aware of the system turn itself, only of the cluster it belongs too. Therefore it may assign a very high score to a candidate that belongs to a very relevant cluster given the dialogue history, but is grammatically incorrect, for example.

Not only does the next utterance accuracy reveal differences between models, but also across languages. There seems to be a big correlation between this accuracy and quality of the pretrained language model. First, the English models outperform the models in lower-resource languages. If we then compare the remaining three languages, Spanish and French are one step ahead of Norwegian. As we will see in the next sections, this will be a recurring phenomenon. The GPT2 model in English was pretrained and released by Open AI. 40 GB of web cleaned and processed data was used. In comparison, we only used 10GB, 7GB and 5GB used to pretrain the Spanish, French and Norwegian models, respectively. Additionally, we also believe that the OSCAR corpus used to increase the amount of data in Norwegian is not as benefitial for our domain as Opensubtitles and Wikipedia, due to its nature. Since it is made of web scrapped text from the Internet, it may contain many sentences that are not related to our task at all, hindering the fine-tuning procedure.

Word overlapping metrics: While the next utterance classification score seem to be very aligned with the expected behaviour of our proposal, the F1 and BLEU score do not seem to be that correlated. Table 5 shows the obtained results. Nonetheless, there are still some conclusions to be made.

Table 5. F1 and BLEU scores obtained by all the models in the four languages in the test partition of the corpus..

	English		Spanish		French		Norwegian	
	F1	BLEU	F1	BLEU	F1	BLEU	F1	BLEU
BL no pretraining	-	-	0.198	0.096	-	-	-	-
BL small	0.303	0.143	-	-	-	-	-	-
BL	0.305	0.139	0.229	0.109	0.250	0.119	0.297	0.147
BL+SC	0.299	0.140	0.248	0.122	0.257	0.124	0.274	0.123
BL+PH	0.283	0.132	0.259	0.130	0.241	0.111	0.297	0.141
FM	0.302	0.143	0.272	0.145	0.259	0.122	0.309	0.149
FM+RR	**0.315**	**0.150**	0.288	0.150	0.275	**0.135**	0.317	0.158
FM+WDH	0.303	0.142	**0.296**	**0.159**	**0.276**	0.131	**0.322**	**0.164**

First of all, the two metrics behave in a very similar manner, which makes sense because both are measures of how the produced system response resembles the ground truth sentence found in the corpus. Second, if we compare the results of the different models, we can see that including the scenario or phase embeddings does not consistently yield to better results. There does not seem to be any difference between the small and medium models in English. On the other hand, not pretraining the baseline in Spanish again yields to worst results. Interestingly, applying a reranking process does improve the result in both metrics across the four languages. This shows that the reranking methodologies play an important role in our system, and that are capable of selecting the responses which are closer to the ones produced by human experts. Finally, we would also like to mention that in this case the results on different languages should not be compared too in depth, because the four languages are morphologically different and therefore the differences might well be due to language particularities instead of to performance discrepancies. In any case, many authors have argued that word overlapping metrics are not highly correlated with the actual quality of the responses [60], because a response that does not share any words with the ground truth reference could indeed be completely appropriate. Thus, we now provide a human evaluation to further validate our proposals.

8.2 Human evaluation of the responses

The quality of the generated responses was measured by coaching experts. Four different experts per language participated in this evaluation. They compared pairs of responses of different models. Per each language and model pair, 40 pairs of responses were ranked twice. Every evaluator assessed the same number of instances per model pair, where each instance consisted of a local dialogue history made of the last 5 turns, and two possible continuations for the system. The dialogue histories were different for each model pair in order not to bias the evaluators. Four options were presented to the evaluators. According to their criteria in the context of coaching and the project, they had to select whether the first response was better, the second one was better, both of them were equally valid to continue the dialogue, or none of them was acceptable. We considered using more fine-grain level metrics, such as the ones used in [105], but we decided to stick to the simpler approach

because: 1) since the evaluators are experts, they should be able to weight the different aspects of the responses and reckon which is more appropriate for the task, 2) the models to be compared should not vary drastically in the style of the responses, because they are different versions of similar methodologies, and 3) it is therefore more cost-efficient; the additional costs would not compensate the potential benefits of a more detailed evaluation, owing to the aforementioned reasons. Additionally, in Section 8.3, we perform a detailed evaluation of the best model, which shows the strengths and weaknesses of our conversational agent in depth.

Table 6 shows the results of the comparison of the models, combined in all the languages. Additionally, Tables 7, 8, 9 and 10 show these results divided by language; in English, Spanish, French and Norwegian, respectively. Binomial tests of significance were carried out in the global comparison of the models shown in Table 6, since it contains more samples and it is therefore more appropriate. The p-value was computed taking into account only the decisive comparisons: *A is better* versus *B is better*.

Table 6. Results of the pairwise response quality evaluation combined in the four target languages. Models in bold indicate that they are significantly better than their counterpart ($p<0.05$).

Model A	Model B	Neither A nor B	A is better	B is better	Both A and B
BL	BL+SC	17.50	26.25	30.00	26.25
BL	BL+PH	18.75	27.81	28.38	24.06
BL	**FM**	17.81	24.38	35.00	22.81
FM	**FM+RR**	20.63	18.75	31.56	29.06
FM+RR	**FM+WDH**	17.50	21.88	31.87	28.75
FM+WDH	**GT**	7.50	19.06	50.94	22.50

Table 7. Results of the pairwise response quality evaluation in English.

Model A	Model B	Neither A nor B	A is better	B is better	Both A and B
BL	BL small	12.50	41.25	17.50	28.75
BL	BL+SC	11.25	33.75	23.75	31.25
BL	BL+PH	12.50	27.50	30.00	30.00
BL	FM	11.25	21.25	32.50	35.00
FM	FM+RR	13.75	13.75	38.75	33.75
FM+RR	FM+WDH	13.75	25.00	32.50	28.75
FM+WDH	GT	10.00	30.00	42.50	17.50

In general, the obtained results are coherent with our proposal and with the automatic evaluation, especially with the next utterance classification accuracy. While only including one of the proposed embeddings to control the dialogue not always results on a better model according to this evaluation, including both significantly improves the quality of the responses compare to the baseline. The effect of the reranking using just the GPT2 score is particularly interesting. Even if, in general, it is significantly better than not using it, there are some difference if we compare the results across languages. It improves the quality of the responses in English the most, followed by French and Spanish. In Norwegian slightly worsens the quality of the responses. This could be closely related to the next utterance classification accuracy, which was shown in Table 4. In English the

Table 8. Results of the pairwise response quality evaluation in Spanish.

Model A	Model B	Neither A nor B	A is better	B is better	Both A and B
BL	BL no pretraining	26.25	37.50	18.75	17.50
BL	BL+SC	13.75	20.00	32.50	33.75
BL	BL+PH	13.75	31.25	30.00	25.00
BL	FM	16.25	23.75	33.75	26.25
FM	FM+RR	18.75	20.00	31.25	30.00
FM+RR	FM+WDH	15.00	26.25	33.75	25.00
FM+WDH	GT	5.00	10.00	52.50	32.50

Table 9. Results of the pairwise response quality evaluation in French.

Model A	Model B	Neither A nor B	A is better	B is better	Both A and B
BL	BL+SC	22.50	27.50	25.00	25.00
BL	BL+PH	23.75	23.75	26.25	26.25
BL	FM	22.50	30.00	36.25	11.25
FM	FM+RR	25.00	13.75	36.25	25.00
FM+RR	FM+WDH	15.00	17.50	31.25	36.25
FM+WDH	GT	3.75	22.50	60.00	13.75

Table 10. Results of the pairwise response quality evaluation in Norwegian.

Model A	Model B	Neither A nor B	A is better	B is better	Both A and B
BL	BL+SC	22.50	23.75	38.75	15.00
BL	BL+PH	25.00	28.75	31.25	15.00
BL	FM	21.25	22.50	37.50	18.75
FM	FM+RR	25.00	27.50	20.0	27.50
FM+RR	FM+WDH	26.25	18.75	30.00	25.00
FM+WDH	GT	11.25	13.75	48.75	26.25

next utterance accuracy is the highest of all languages, and therefore the model selects candidates with are often closer to what a human would select. Then French and Spanish are next, and so their improvement is not as big as in the English model in this case. Finally, the worst accuracy is obtained in Norwegian, which may well indicate that the GPT2 score by itself is not reliable to successfully select good candidates. Moreover, if we now focus on the influence of adding the WDH score instead of using only the GPT2 score, we can see that it consistently improves the quality of the responses. This definitely makes sense since it already showed an improvement in terms of next utterance accuracy, as shown in Figure 5. However, it is important to remark that in this case the reranking is carried out over a set of model-generated candidates, while in the previous study of the next utterance accuracy the candidates where human responses from the corpus. This indicates that the WDH system is robust no matter the nature of the candidates. Finally, our full proposal (FM+WDH) was compared with the ground truth responses of the corpus. As expected, the ground truth significantly outperforms our model in all the languages. However, the margin in

English is remarkably small, which shows that a better pretraining is key to develop end-to-end dialogue models.

In this regard, an additional comparison was carried out to measure the effect of not pretraining the baseline model in Spanish (first row in Table 8). It underlines the fact that a pretrained language model is essential to enhance the posterior performance of the dialogue model. A similar study was carried out in English. In this case, we compared the *small* and *medium* pretrained GPT2 architectures (first row in Table 7). The medium architecture showed its superiority, as it has done in many other NLP tasks [80]. We were not able to pretrain medium models for the other languages due to the lack of computational and corpus resources.

8.3 Human interaction evaluation

Let us introduce the results of the human interaction with the FM+WDH system. The same four evaluators per language that carried out the evaluation of the responses were the ones interacting with the system. Additionally, some of the non-English evaluators but fluent in English also tested the English system. Thus, the English system was evaluated by 12 experts, and the rest of the systems by 4. Each evaluator carried out two dialogues with the corresponding system, first the introductory dialogue into coaching, and afterwards the first part of a GROW nutrition session. On average, the dialogues were 40 turns long (20 user turns + 20 system turns). This value was controlled via the dialogue phase embeddings. After interacting in the two scenarios, the evaluators filled the aforementioned CUQ and HFQ questionnaires (Appendix C). Table 11 shows the mean and standard deviation of the score achieved in these two questionnaires, divided per language.

Table 11. CUQ and HFQ mean scores (and standard deviation), divided per language.

Language	CUQ score	HFQ score
English	69.1 (12.9)	63.1 (18.0)
Spanish	62.1 (11.0)	62.5 (13.8)
French	68.7 (6.6)	61.9 (11.2)
Norwegian	39.1 (11.1)	43.8 (11.8)

The English, Spanish and French models achieved a score higher than 50 in both tests, which means that on average the evaluators tended to agree on the positive aspects of the system and disagree on the negative ones. On the contrary, this was not the case for the Norwegian system, which shows that there is still a significant performance gap to be closed for systems in languages with very few resources. The English model achieved the best results once again, but interestingly enough, the French and Spanish models were unexpectedly close in terms of HFQ score, and the French one was very close in the CUQ score too. This might be due to the fact that the pretraining of the GPT2 models affects mostly on the candidate generation stage, whereas the WDH system is only (except the turn embeddings) learnt on our coaching corpus. The WDH reranking is a key aspect on the whole pipeline, since it is the main responsible of keeping coherence in the dialogue, which greatly influence the user experience. This is even more important when dealing with coaching dialogues where the long-term strategy is so valuable. Thus, we hypothesize that future improvements in this direction would result in more structured, and therefore better rated, dialogues. In the case of the Norwegian, however, the main issue might be that, overall, the generated candidates lack quality due to the worse pretraining of the GPT2 generative model. If this were the case, improving the quality of multilingual transformers or a better pretraining in low resource

languages would be essential to improve the usability and emotional influence of this kind of models in the future.

Let us now focus on the specific answers of each questionnaire. Figures 6 and 7 shows the average results and their standard deviation for each question in the CUQ and HFQ, respectively. The values have been computed with the combination of the questionnaires in the four target languages.

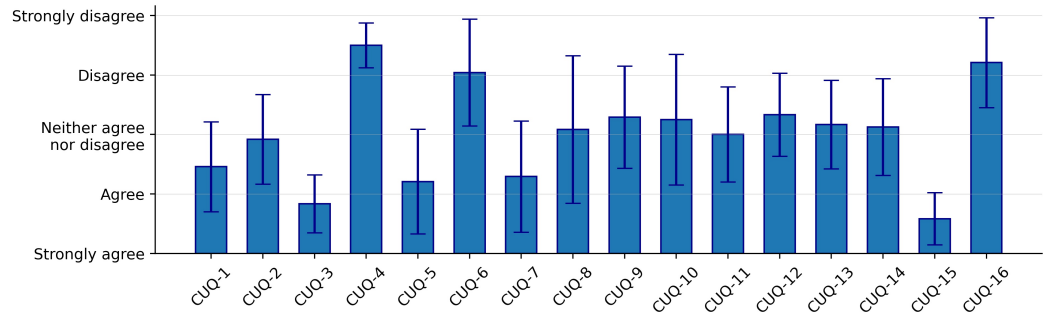


Fig. 6. Results of the Chatbot Usability Questionnaire.

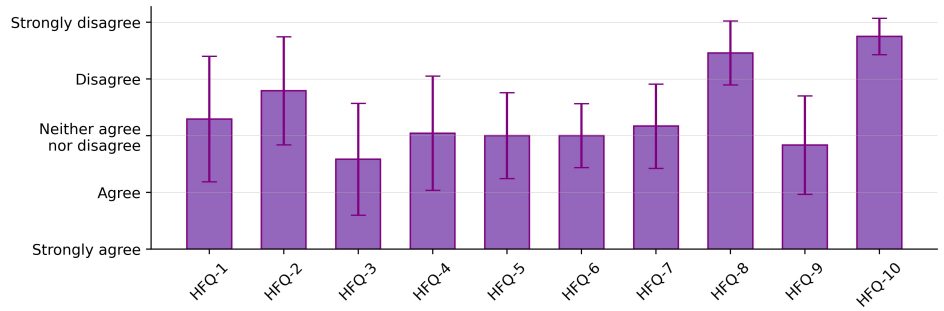


Fig. 7. Results of the Hedonic Feelings Questionnaire.

As for the usability, the responses to questions ranging from CUQ-3 to CUQ-6 indicate that the dialogue system presents itself correctly and that indicates well its purpose. More exactly, this means that the user understands that a session of coaching is about to be carried out, and that to this end they will first talk about the user's hobbies before getting into the nutrition GROW session. Responses to CUQ-7, CUQ-8, CUQ-15 and CUQ-16 indicate that the interaction with the system is rather simple and easy, which is an important point for future interactions with real users. In general, the performance is not that great in terms of understanding the user and acting accordingly, as reflected in the results of questions from CUQ-9 to CUQ-12. In this regard, it is important to recall that the proposed methodology does not make use of any explicit knowledge representation like entities, ontologies or dialogue acts. It purely learns from the text transcription of dialogues. This makes possible to develop a dialogue system in an easier and more affordable way, but it also has its limitations. The systems is less likely to react to user turns that contain some relevant information than if a Natural Language Understanding module was used, for example. CUQ-13 and CUQ-14 refer to the ability of the system to recover from errors. It seems that the system can recover from errors sometimes, but that other times it fails to do so. This is definitely

an interesting and open topic of research, and we plan to use the WDH system to detect dialogue breakdowns, and avoid them if possible. Finally, responses to CUQ-1 and CUQ-2 indicate that the system is engaging to some degree, but that it is also quite robotic.

On the other hand, the HFQ provides useful information to measure the potential impact the system may have on the user, at least short-term. Very importantly, experts strongly agree that the interaction with the system is neither depressing nor stressful (HFQ-8 and HFQ-10). This is a good starting point, because at least the system does not seem to give rise to very negative feelings. It does not seem to be boring either (HFQ-2). On the other hand, the HFQ also reveals that there is much progress to be done, since the system could be much more stimulant (HFQ-7). Coaching is about stimulating the user in order to help them to achieve their own goals. Thus we would really like to improve in this aspect. Nonetheless, experts do not think the communication is not stimulant either, which also means are not completely away from our objective. Apart from this, experts feel the system is quite innovative (HFQ-3) and do not agree nor disagree on the fact that the communication with system is extraordinary (HFQ-1), disappointing (HFQ-4), thrilling (HFQ-5), trivial (HFQ-6) or reassuring (HFQ-9). Being able to produce dialogues even more coherent long-term would likely result on improvements on these aspects.

In summary, these result indicate that, in general, our proposals are heading in the right direction, but also that improvements are probably needed to systematically use our coaching system with end users.

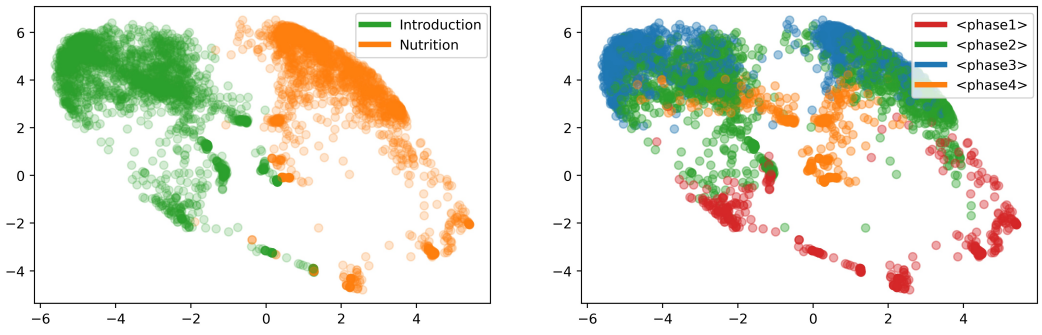
9 THE WDH SYSTEM AS A TOOL TO EXPLAIN THE BEHAVIOUR OF THE CONVERSATIONAL AGENT

The WDH system has shown to improve the response quality of the system by integrating the whole dialogue history into the decision making stage. Additionally, it can also be a powerful tool understand on what basis these decisions have been taken. In this section we will first analyse the distribution of turn embeddings in the low dimensional space. This can help us understand how the turns are clustered, and intuitively validate those, also by comparing them to dialogue acts. Additionally, we will arrange the clusters and dialogue acts into graphs to visualize the paths the system is more likely to take and understand why. Moreover, we believe this kind of analysis could be taken one step beyond, and use it not only to analyse but also to improve the behaviour of the system. We leave this interesting research topic for future work.

9.1 Low dimensional turn embedding space

In all the presented experiments the low-dimensional turn embedding space has been of size 5. Empirically, it has been a good choice to provide interesting results and to make the WDH work. However, we can also choose to convert the high dimensional turn embeddings into bidimensional, and therefore visualizable, vectors. While this can be done by projecting the 5 dimensional vectors into two dimensions with another dimension reduction technique, we have opted to train a second supervised autoencoder. We believe that this way the distribution of the points (system turns) in the bidimensional space should be more similar to the one in the 5 dimensional one.

For example, this way we can see the clusters the turns are grouped in. We have shown this distribution back in Section 4, in Figure 1. There, the number of clusters is 20, lower than the actual 60 used in our experiments, for the sake of clarity. They correspond to the English version of the corpus. In that figure some turns that have been grouped together have been highlighted. This manual inspection already suggests that turns clustered together share semantic information. Additionally, much more patterns can easily be detected. For instance, we can also group the system turns according to the scenario or dialogue phase they belong, as shown in Figure 8.



(a) Scenario.

(b) Dialogue phase.

Fig. 8. Bidimensional turn embeddings grouped by the scenario and dialogue phase they belong to.

The groups in this case make a lot of sense. For instance, we can see that the systems turns corresponding to the two scenarios are well separated. Nonetheless, there are two areas where these are much closer. If we check the dialogue phase distribution, we see that these areas correspond to the first and last dialogue phases. This seems very much reasonable, because the greetings and the goodbyes are similar in both scenarios, or at least much less different than the rest of the system turns.

There are also other properties of the corpus that become visible in this space. For example, we can take advantage of our corpus being labeled [99], even if we have not been used this labels at any stage of the development of the dialogue system. We can check the distribution of the turns according to their labels. This is shown in Figure 9.

For sake of simplicity, we are only showing the turns corresponding to a subset of the labels, and some labels have been merged for a better visualization (some different types of questions about nutrition are merged into just *Nutrition question*, for example). For example, Figure 9 shows that the turns labeled as *Hello* are placed in the same place as the ones corresponding to the first dialogue phase, and the same applies for the *Goodbye* and the last dialogue phase turns. Turns labeled as *System introduction* occupy the same space than the introductory scenario turns of the first and second dialogue phase, suggesting that, in fact, the system presents itself at the beginning of the first scenario. *Travelling* and *Music/hobbies* are very close in the bidimensional space, roughly in the place of the second and third phase of the introductory scenario. These are actually two of the topics the system usually covers to make the user feel more comfortable.

In the right hand side of the space we can find the turns categorized as *Objective* and *Nutrition question*, which clearly correspond to the GROW session about nutrition. They seem to be in a very similar region. This could be due to the dimension reduction being too drastic, and they could perfectly be better separated in a higher dimensional space. Additionally, the *Nutrition question* turns seem to be a bit more widespread than the *Objective*. This is coherent with the GROW coaching strategy: the system asks questions about nutrition in many situations, but only focuses on the objective once the user has confirmed that there is in fact something they would like to achieve.

There are also turns labeled as *Confirmation*, which are to be found all over the place in both scenarios. This definitely makes sense; the system may have to confirm whatever the user asks at any point. Finally, we have *Change topic* turns, which are located in both scenarios but only in a

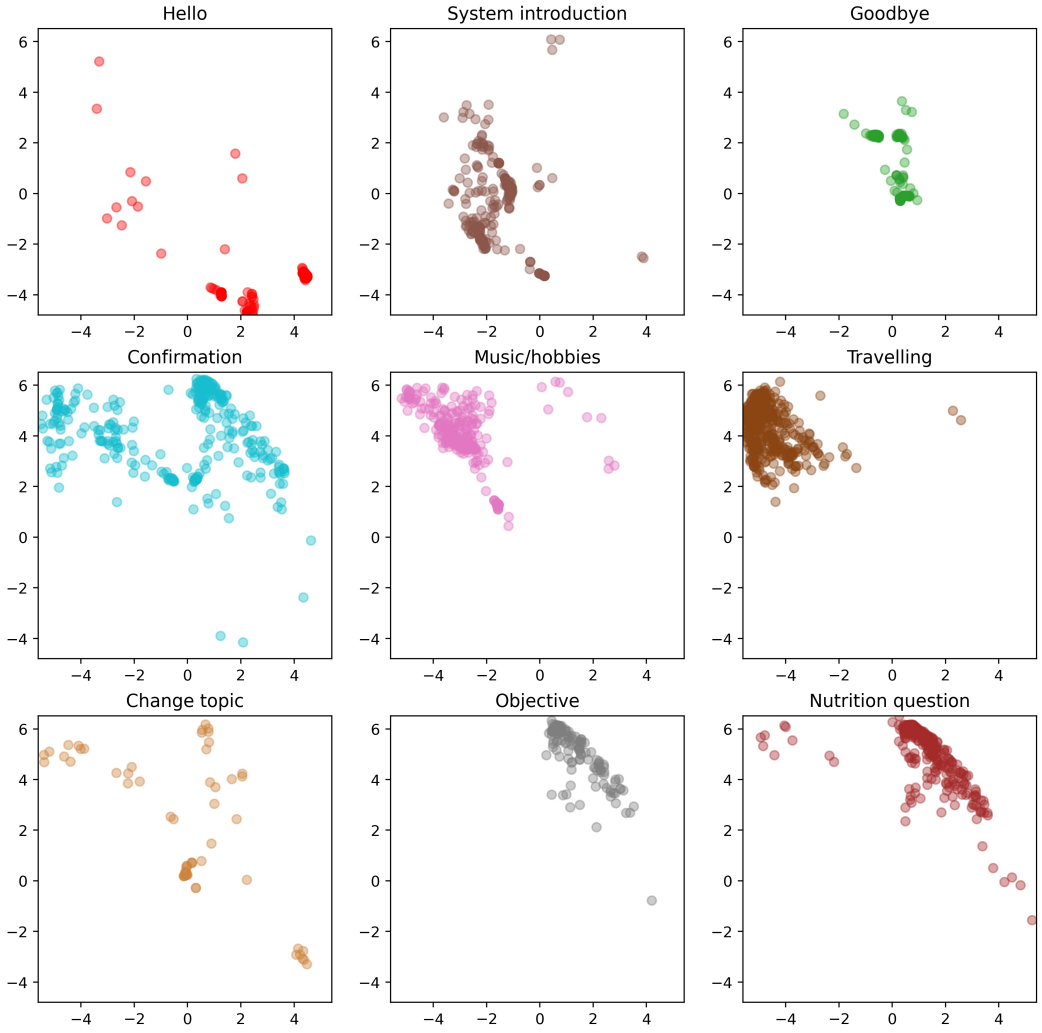


Fig. 9. Bidimensional turn embeddings divided according to the dialogue act labels.

few areas. These correspond to utterances where the system and the user finish talking about a given topic, and the system or the user suggest a new one.

9.2 Clustering as an unsupervised way of learning dialogue acts

The fact that this low dimensional turn embedding space is so structured also validates the proposed methodology from a more intuitive point of view: if the turns that are close in the low dimensional space share semantic information, can often be labeled with the same dialogue act and are used in similar situations in the dialogue, then the resulting clusters should also represent that information. Therefore, might clustering be considered an unsupervised way of learning dialogue acts to a certain extent? To answer this question, we perform dialogue act classifications from turn embeddings and from cluster indexes. If clusters act as unsupervised dialogue acts, both classifications should produce similar or at least comparable results.

To this end, we employed a bigger set of dialogue acts than the shown in Figure 9. For example, the *Nutrition question* label was subdivided *Motivational question*, *Resources or Obstacles question*, and so on. As a result, a set of 26 dialogue acts were finally used as the classification targets. These are listed and described in Appendix D. We perform three series of experiments. First, we attempt the dialogue act classification task from turn embeddings, via a simple two-layer feed forward neural network. Second, we do the same, but from the low dimensional embeddings, which were as the input to the clustering method in order to avoid the curse of dimensionality issue, as mentioned back in Section 6.2. Thus the comparison might be more fair, since the clustering module and the classifier will have exactly the same input. Third, we try to predict the dialogue act only from the cluster a system turn has been assigned to. To do so, we learn a mapping from cluster indexes to dialogue acts in the training partition of the corpus, applying a (multi-start) local search heuristic optimization to maximize the F1 score. Specifically, a first improvement heuristic was employed, and two mappings were considered neighbors if and only if they only differed in one value, i.e. if one and only one cluster was mapped to a different dialogue act. Table 12 shows the F1 scores of the three classification methods in the test partition of the corpus in the four target languages.

Table 12. F1 scores of the three classification methods in the test partition of the corpus in the four target languages.

F1 Score at dialogue act classification	English	Spanish	French	Norwegian
From turn embeddings	0.505	0.498	0.483	0.473
From dimensionally reduced turn embeddings	0.328	0.293	0.299	0.292
From cluster index	0.287	0.279	0.285	0.285

In general terms, the F1 scores are reasonable, considering that this challenging task involves a classification between 26 quite imbalanced classes, where the majority class is around 26.4 more frequent than the minority class. Mind that a random classifier obtains an F1 score of around 0.03. It can be seen that the results follow the same pattern across the four languages. The best results are achieved, as expected, with the whole turn embeddings. Then the F1 score drops around two tenths if the lower dimensional embeddings are used. However, interestingly, the difference between the classification from the low dimensional turn embeddings and from the cluster indexes is rather marginal. This comparison is very important, since both algorithms takes as input the same dimensionally reduced embeddings. Therefore, it seems that clustering is able to extract almost the same information about dialogue acts from those embeddings than a classifier trained specifically to do so. Thus, it seems reasonable to say that, indeed, clustering works as an unsupervised way of learning dialogue acts. At least, there seems to be a strong correlation between the learnt clusters and the dialogue acts.

In order to gain a deeper insight into this correlation, it is specially interesting to analyze how the F1 score of the cluster to dialogue mapping changes with respect to the number of clusters. This is shown in the Figure 14 in Appendix B, it is the purple line in the plot. The F1 score grows a lot from 10 clusters to 50, but from 60 clusters on it stabilizes. Thus, the optimal number of clusters (around 60), is quite higher than the number of dialogue acts, the double approximately. This indicates that if the number of clusters is too low, some of the clusters will contain turns with many different dialogue acts. After splitting them, when the optimal number of clusters is reached, there will be multiple clusters mapped to the same dialogue act. The turns within these clusters will probably differ in the context they are used: since the low-dimensional turn embeddings are learnt in a way that they contain information about the scenario and the dialogue phase, it might happen that the

clustering makes some distinctions where the dialogue acts do not. For example, if we consider the system turns "*I understand that you have a healthy eating routine*" and "*I understand, you really love travelling*", it may perfectly happen that they are assigned to different clusters, one that contains mainly similar sentences about nutrition and the other one into a cluster more related to travelling or to the introductory scenario. However, regarding the dialogue act, both would be labeled as *I understand*. Last, if we increase the number of clusters even more, the F1 score does not notably change anymore. This is probably due to some clusters being divided, but then they being mapped to the same dialogue act. Thus, the classification results are very similar.

9.3 Cluster and dialogue act dynamics

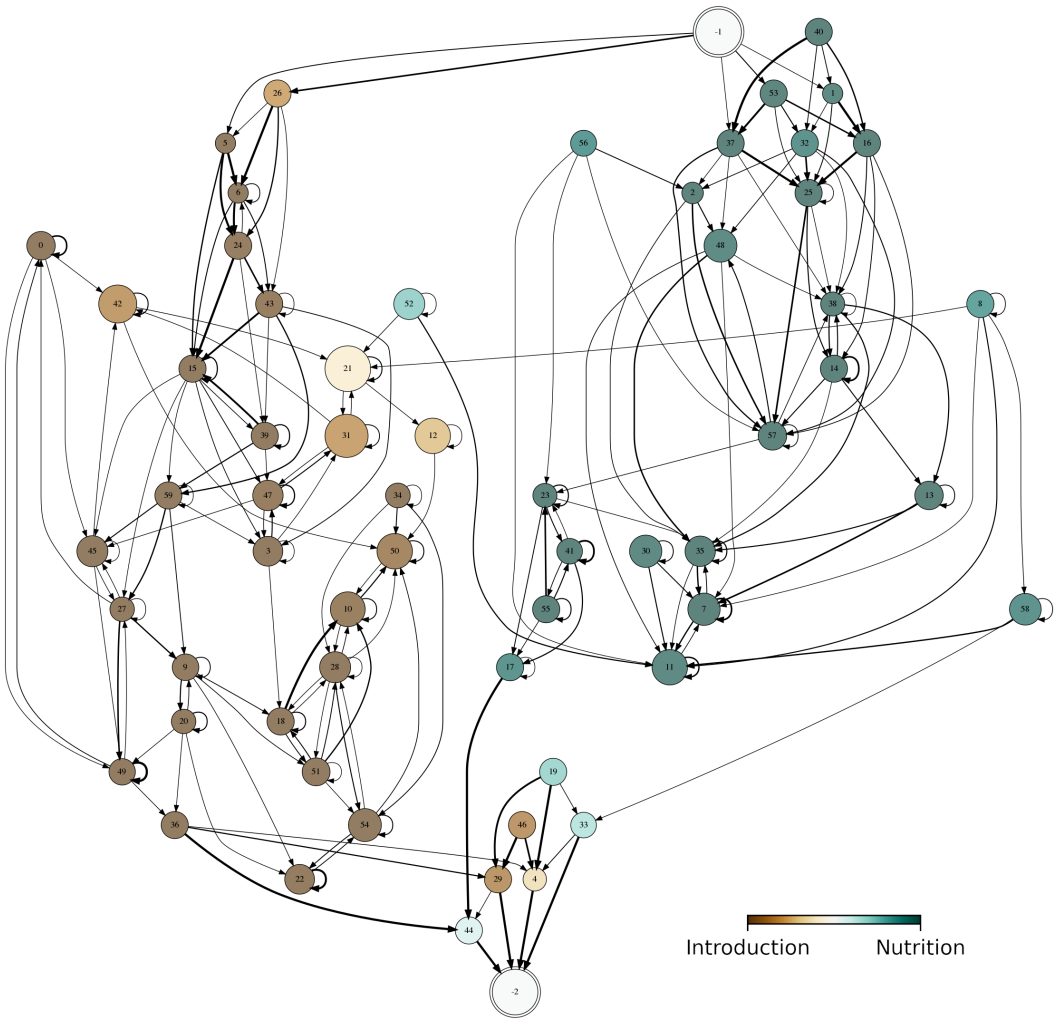


Fig. 10. A graph where nodes represent clusters, and their colors the scenario of the turns they gather.

This relation between dialogue acts and clusters is also visible if we analyse the dialogue flow. This can be done by arranging the clusters or dialogue acts into a graph that shows the number of

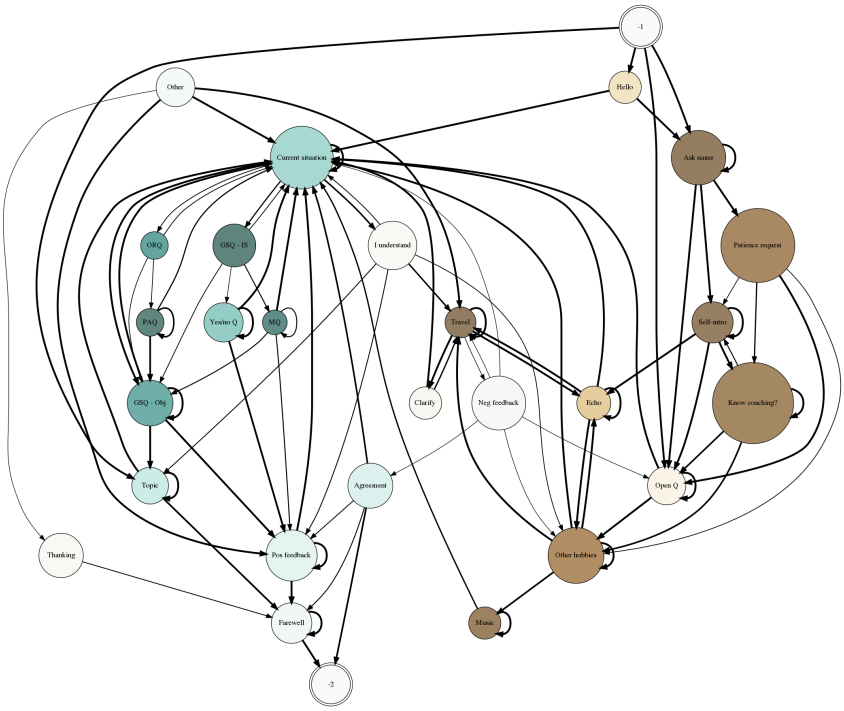


Fig. 11. A graph of dialogue acts, coloured according to the scenario they were used in.

transitions between each of them. We show such graphs for clusters in Figures 10 and 12, and for dialogue acts in Figures 11 and 13. These have been built with the English version of the corpus. In all the diagrams, the node -1 is the source, i.e. the nodes that come after it represent the cluster/act of the first system utterance in dialogue. On the other side, the node -2 is a sink; it represents the end of a dialogue. To keep the graphs as informative as possible, we skip some minor transitions: we do not show edges that correspond to less than the five percent of the total transitions from a cluster/act to another. This is the reason why some nodes have no edges in their direction in the figures.

In Figure 10, the clusters have been coloured according to the scenario of the turns they gather. The browner nodes refer to clusters that mainly contain turns used in the introductory scenario. Alternatively, the greener ones correspond to clusters related with the GROW session about nutrition. The same color scheme has been applied in Figure 11, but for dialogue acts. If we focus on Figure 10, it is interesting that, while most of the clusters are one-sided, there are some few that share introductory and nutrition turns. These often include generic turns like confirmations or backchannels. In general, we can see that the graph can be split into two major regions: the browner one that corresponds to introductory dialogues, and a greener one which unravels the structure of the coaching sessions about nutrition. The two regions merge almost exclusively at the end of the dialogues, when the system bids farewell to the user. However, the dialogue act graph in Figure 11 is not so split. Even though there are many dialogue acts that clearly correspond to one scenario, many others are colored in white, such as, *Thanking*, *I understand*, *Neg. feedback* or *Clarify*. As aforementioned, the clusters corresponding to these acts have probably been broken into several different clusters. Another dialogue act that is probably divided into many clusters is

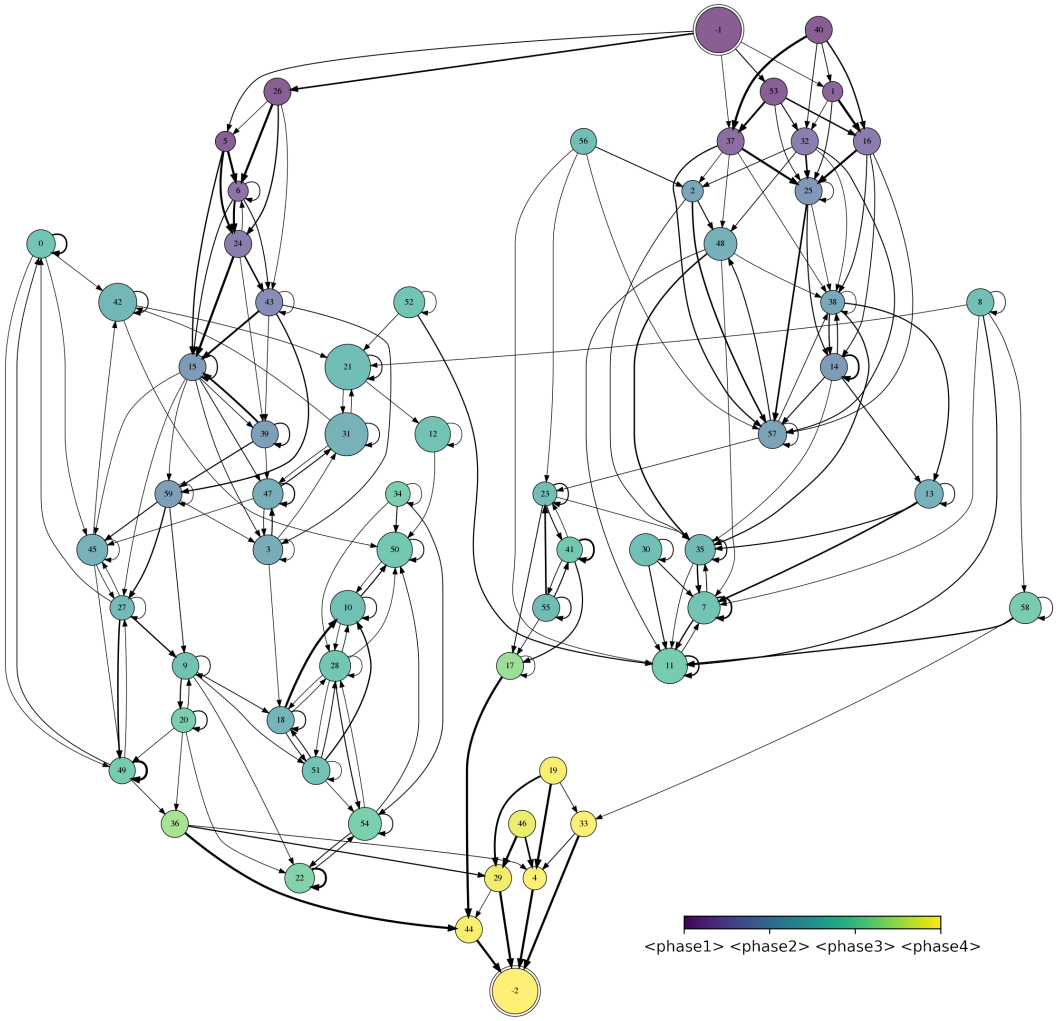


Fig. 12. A graph where nodes represent clusters, and their colors the dialogue phase of the turns they gather.

Current situation, which is a very central node in the graph, meaning that it is used in different contexts. This makes sense, since it is necessary to analyse the user's current situation in order to establish a goal and carry out the coaching session accordingly.

On the other hand, the graphs shown in Figures 12 and 13, where the nodes are coloured in terms of the dialogue phase, show similar patterns. For example, in this case we can deduce that the *Topic* label has also been divided into multiple clusters. On the one hand it is colored in blue/green which means that it contains many turns used when the dialogue is quite advanced; but there is also a an arrow from -1 to it, denoting that there are many dialogues that start with that dialogue act. This really makes sense, because the *Topic* dialogue act groups utterances that open, close or choose a new topic. This distinction has probably been learnt in the clustering, but is not shown in the dialogue act graph.

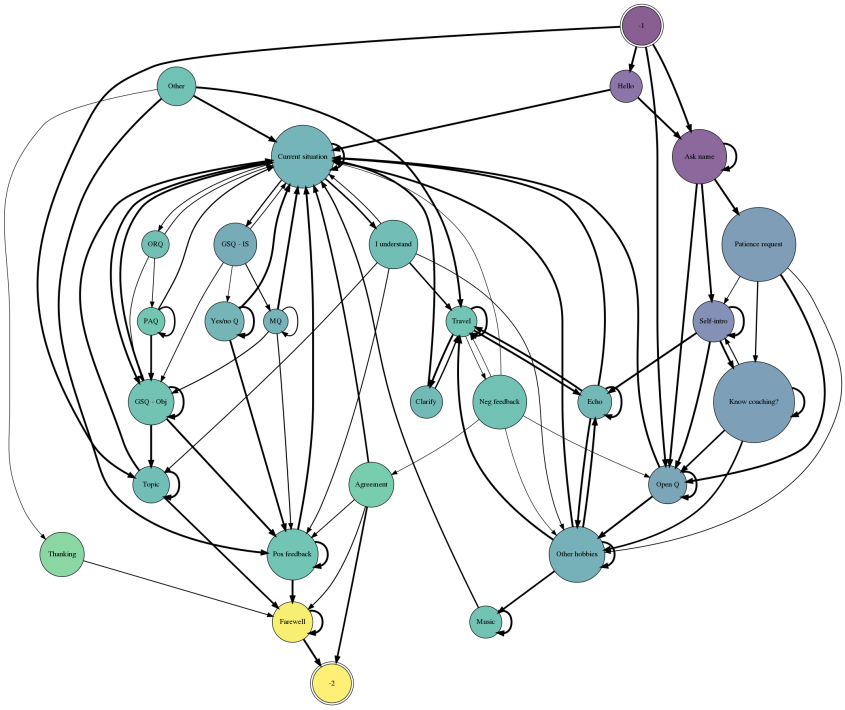


Fig. 13. A graph of dialogue acts, coloured according to the dialogue phase they were used in.

10 DISCUSSION OF FINDINGS AND IMPLICATIONS

Let us summarize the most notable findings of our research and their implications on developing intelligent conversational agents.

Bridging the gap between state-of-the-art Artificial Intelligence techniques and current coaching models. If we compare the dialogue technologies used in coaching agents found in the literature and the market and the ones employed in the most novel and prominent chatbots, there is a big disparity. This is valid for most of the health-care related conversational systems too. In a nutshell, professional dialogue strategies in health-care related conversational agents are often implemented, at least partially, via hand-crafted policies. On the other hand, state-of-the-art dialogue models are fully data driven, and thus do not require carefully designed policies, these are learnt from the data. In this work we have shown that it is possible to adapt and modify these novel technologies to develop complex coaching conversations. This provides mayor benefits. First, explicit expert knowledge does not need to be provided to develop such conversational agents, which definitely simplifies the whole process of building them. Second, general purpose neural dialogue models such as GPT2, which form the base of our conversational agent, have shown extraordinary abilities to learn and generalize for many tasks. Thus, the resulting dialogue models might potentially perform better than rule-based models, which can only work for a limited amount of domains and situations. Nonetheless, there are still limitations to this attractive approach. Its main drawback is a consequence of the models being fully data driven. It may happen that, for instance, the system makes an error at some point in the dialogue, as a result of a non-completely successful training; or there might also be some inconsistencies with the name entities, because the system is not able to automatically coherently keep track of them. While those errors could be solved in a rule-based

system easily, they have no direct solution in a fully data driven model. Our proposals help alleviate this issue by enhancing coherent responses, but they do not ensure errors will not happen. However, we believe that we are in the point where the aforementioned benefits start to compensate these limitations; where the ease of building the systems and their better and more general responses make some minor errors affordable. Thus, we foresee that more and more coaching agents, and intelligent conversational agents in general, will adapt methodologies similar to the ones presented here.

These methodologies are mostly ready for multilingual conversational systems. Not only do we reckon that the time for end-to-end conversational agents has already arrived, but we also think that this holds for models in many languages other than English. Additionally, the fact that our models in English, Spanish, French and Norwegian require just the same engineering effort is also encouraging. In contrast, classical modular dialogue systems require the development of some very language dependent modules, such as the Natural Language Understanding or Natural Language Generation modules, which often multiply the effort needed to develop the conversational system in an additional language. In our case, the only limitation is the pretraining step of the generative model. In this regard, we have shown that pretraining the GPT2 model on languages like French or Spanish with open domain corpora such as OpenSubtitles or Wikipedia leads to only slightly worse results than the ones obtained with the official pretrained English model. However, the experts evaluated the Norwegian system much poorer, due to less data being available for the pretraining. In any case, with more and more research targeting non-English languages, we believe that the difference in performance of conversational agents and other NLP models in English and in other languages will rapidly attenuate in the near future, and that therefore language-agnostic approaches like ours will gain popularity and perform even better.

Our work and proposals are general; they can be applied in many other tasks and contexts. We will now discuss our two main methodological contributions, but first we would like to remark that, even though they aim at improving the performance and coherence of a coaching model, they are also valid when developing end-to-end dialogue systems for many other applications. On the one hand, the scenario embeddings allow generative models to select which scenario or they should carry out, which can come out handy in any multi-task set up. On the other hand, the dialogue phase embeddings and the WDH system permit to model and improve the long-term structure of the dialogue. This can be beneficial for almost every dialogue system, from task-oriented to even open-domain, since all of them need to be coherent long-term. In fact, open-domain dialogues or chit-chatting have traditionally been treated as completely unstructured, but there are always underlying phases, even in open social dialogues [32].

Improved response generation by conditioning the generative network using scenario and dialogue phase embeddings. Our first methodological contribution are the scenario and dialogue phase embeddings. Learning these kind of embeddings are very simple and very flexible too. As aforementioned, scenario embeddings could be used in multi-task or multi-domain environments, which have recently gained a lot of interest from the dialogue community [25, 82]. Apart from enabling the use of a single system for all the domains, the learnt embeddings could also provide information about each task and serve as a tool for comparing them. We have not carried out such analysis in this work because there are only two domains in our corpus, but studies in other corpora could be a direction for future work. As for the dialogue phase embeddings, for the moment we have predefined when a dialogue phase starts and when it ends, based on a manual inspection of our data. However, we believe that this approach could be further enhanced, probably with mechanisms that learn the beginning and the end of a phase in a unsupervised way. We think that the WDH system could be useful to this end.

Improved long-term coherence via the WDH system and unsupervised dialogue act learning. The proposed WDH system has shown a great potential. It has improved the performance of our baseline models in automatic and human evaluation in the four target languages, showing that dialogue models actually require long-term context information to keep more coherent conversations. Not only that, we have also analysed the clustering process inside this long-term context system, and a strong correlation with dialogue acts has been found. More precisely, as the experiments carried out in Section 9.2 indicate, the clusters the system turns have been grouped in share to a certain extent the dialogue act they were assigned in a manual labeling. In other words, clustering system turns and then mapping the corresponding cluster into a dialogue act is almost as effective as directly applying supervised learning from the low-dimensional turn embeddings, and not exceedingly worse than classifying the whole turn embedding. Thus, we hypothesize that building a similar system that uses dialogue acts would not outperform our proposal by a big margin. This is a big deal, since many conversational agents rely on dialogue act representations, which involve costly and time-consuming annotations. We hope that our efforts to find alternatives will trigger other researchers' interest on alternative (and potentially unsupervised) turn representations, which could simplify the process of building and designing conversational systems. On the other hand, the main downside of not employing explicit turn representations like dialogue acts or name entities is the potential external control over the system is reduced, i.e., it is harder to explicitly manipulate the behaviour of the system in the case this was necessary in some application. We plan to research on this subject in the future, with the goal that end-to-end conversational agents will be more reliable and robust in the future. We might investigate methodologies that make use of the proposed WDH system to this end. We have already shown that it can be used to visualize many patterns that appear in our data and to better understand why the system might prefer some candidates over others. We hypothesize that there might be some way to use this information explicitly to tweak the dialogue in some way.

11 CONCLUSION

We have presented an end-to-end neural coaching model capable of learning to carry out coaching sessions in English, Spanish, French and Norwegian. Automatic and human evaluation have shown that our proposals to enhance the long-term structure improve the baseline model in all the languages. The evaluation of the experts' interaction with the system reveals that its performance is rather good in terms of usability and the emotions it may provoke on the users. However, this only holds for the English, Spanish and French systems. In the case of the Norwegian one, the results are much poorer, to the worse pretraining of that model. Additionally, even though the evaluations in the rest of the languages show acceptable results, this does not mean that our system is perfectly ready to be used in real environments. The system still makes some errors and does not always provide completely coherent responses. We plan to investigate how to use the turn representations learnt by the clustering in the WDH system to explicitly correct some of the most consistent errors. In any case, we hope that our ideas can be helpful for other researches and encourage them to apply state-of-the-art dialogue modelling techniques in healthcare and well-being. Moreover, since our proposals are general and can be applied in most dialogue tasks, we also expect that many others can benefit from our work, and adopt similar approaches to improve the behaviour of their dialogue systems.

ACKNOWLEDGMENTS

This work has been partially funded by the Basque Government under grant PRE_2017_1_0357 and by the European Union's Horizon 2020 research and innovation programme under grant agreement No 769872.

REFERENCES

[1] David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. 1985. A learning algorithm for Boltzmann machines. *Cognitive science* 9, 1 (1985), 147–169.

[2] Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *Computing Research Repository* arXiv:2001.09977 (2020). https://arxiv.org/abs/2001.09977

[3] Icek Ajzen et al. 1991. The theory of planned behavior. *Organizational behavior and human decision processes* 50, 2 (1991), 179–211.

[4] Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. 1999. OPTICS: ordering points to identify the clustering structure. *ACM Sigmod record* 28, 2 (1999), 49–60.

[5] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2019. A simple but tough-to-beat baseline for sentence embeddings. In *5th International Conference on Learning Representations, ICLR 2017*.

[6] Ali Orkan Bayer, Evgeny A Stepanov, and Giuseppe Riccardi. 2017. Towards End-to-End Spoken Dialogue Systems with Turn Embeddings.. In *INTERSPEECH*. 2516–2520.

[7] RV Belfin, AJ Shobana, Megha Manilal, Ashly Ann Mathew, and Blessy Babu. 2019. A Graph Based Chatbot for Cancer Patients. In *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*. IEEE, 717–721.

[8] Richard Bellman. 1966. Dynamic programming. *Science* 153, 3731 (1966), 34–37.

[9] Robbert Jan Beun, Siska Fitrianie, Fiemke Griffioen-Both, Sandor Spruit, Corine Horsch, Jaap Lancee, and Willem-Paul Brinkman. 2017. Talk and Tools: the best of both worlds in mobile user interfaces for E-coaching. *Personal and ubiquitous computing* 21, 4 (2017), 661–674.

[10] Hemanthage S. Bhathiya and Uthayasanker Thayasivam. 2020. Meta Learning for Few-Shot Joint Intent Detection and Slot-Filling. In *Proceedings of the 2020 5th International Conference on Machine Learning Technologies* (Beijing, China) *(ICMLT 2020)*. Association for Computing Machinery, New York, NY, USA, 86–92. https://doi.org/10.1145/3409073.3409090

[11] Timothy W Bickmore, Daniel Schulman, and Candace Sidner. 2013. Automated interventions for multiple health behaviors using conversational agents. *Patient education and counseling* 92, 2 (2013), 142–148.

[12] John Brooke. 1996. SUS: a "quick and dirty'usability. *Usability evaluation in industry* (1996), 189.

[13] Paweł Budzianowski and Ivan Vulic. 2019. Hello, It's GPT-2-How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems. *EMNLP-IJCNLP 2019* (2019), 15.

[14] Zoraida Callejas, David Griol, Kawtar Benghazi, Manuel Noguera, Gerard Chollet, María Inés Torres, and Anna Esposito. 2020. Measuring and Fostering Engagement with Mental Health E-Coaches. In *Companion Publication of the 2020 International Conference on Multimodal Interaction* (Virtual Event, Netherlands) *(ICMI '20 Companion)*. Association for Computing Machinery, New York, NY, USA, 275–279. https://doi.org/10.1145/3395035.3425316

[15] Andrew D Carlo, Reza Hosseini Ghomi, Brenna N Renn, Michael A Strong, and Patricia A Areán. 2020. Assessment of real-world use of behavioral health mobile applications by a novel stickiness metric. *JAMA network open* 3, 8 (2020), e2011978–e2011978.

[16] Jacky Casas, Elena Mugellini, and Omar Abou Khaled. 2018. Food diary coaching chatbot. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. 1676–1680.

[17] Chen Chen, Lisong Qiu, Zhenxin Fu, Junfei Liu, and Rui Yan. 2019. Multilingual Dialogue Generation with Shared-Private Memory. In *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 42–54.

[18] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[19] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300705

[20] Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*. 7059–7069.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[22] Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition*. Springer, 187–208.

[23] Fabien Dubosson, Roger Schaer, Roland Savioz, and Michael Schumacher. 2017. Going beyond the relapse peak on social network smoking cessation programmes: ChatBot opportunities. *Swiss medical informatics* 33, 00 (2017).

[24] Charles Elkan. 2003. Using the triangle inequality to accelerate k-means. In *Proceedings of the 20th international conference on Machine Learning (ICML-03)*. 147–153.

[25] Mihail Eric, Rahul Goel, Shachi Paul, Adarsh Kumar, Abhishek Sethi, Peter Ku, Anuj Kumar Goyal, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tur. 2019. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. *arXiv preprint arXiv:1907.01669* (2019).

[26] Anna Esposito, Terry Amorese, Marialucia Cuciniello, Antonietta M Esposito, Alda Troncone, Maria Inés Torres, Stephan Schlögl, and Gennaro Cordasco. 2018. Seniors' acceptance of virtual humanoid agents. In *Italian forum of ambient assisted living*. Springer, 429–443.

[27] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (Portland, Oregon) *(KDD'96)*. AAAI Press, 226–231.

[28] Ahmed Fadhil, Gianluca Schiavo, and Yunlong Wang. 2019. CoachAI: A Conversational Agent Assisted Health Coaching Platform. *arXiv preprint arXiv:1904.11961* (2019).

[29] Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. *Computing Research Repository* arXiv:1707.02633 (2017). https://arxiv.org/abs/1707.02633

[30] Xiang Gao, Yizhe Zhang, Sungjin Lee, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2019. Structuring Latent Spaces for Stylized Response Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 1814–1823.

[31] Aditya Gaydhani, Raymond Finzel, Sheena Dufresne, Maria Gini, and Serguei Pakhomov. 2020. Conversational Agent for Daily Living Assessment Coaching. In *CEUR Workshop Proceedings*, Vol. 2760. CEUR-WS, 8–13.

[32] Emer Gilmartin, Christian Saam, Carl Vogel, Nick Campbell, and Vincent Wade. 2018. Just Talking - Modelling Casual Conversation. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, Melbourne, Australia, 51–59. https://doi.org/10.18653/v1/W18-5006

[33] Rahul Goel, Shachi Paul, and Dilek Hakkani-Tür. 2019. HyST: A Hybrid Approach for Flexible and Accurate Dialogue State Tracking. *Proc. Interspeech 2019* (2019), 1458–1462.

[34] Sergey Golovanov, Rauf Kurbanov, Sergey Nikolenko, Kyryl Truskovskyi, Alexander Tselousov, and Thomas Wolf. 2019. Large-scale transfer learning for natural language generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 6053–6058.

[35] Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D'Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, et al. 2020. Overview of the ninth dialog system technology challenge: Dstc9. *arXiv preprint arXiv:2011.06486* (2020).

[36] Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. 2020. End-to-End Neural Pipeline for Goal-Oriented Dialogue Systems using GPT-2. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 583–592. https://doi.org/10.18653/v1/2020.acl-main.54

[37] Marc Hassenzahl, Michael Burmester, and Franz Koller. 2003. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In *Mensch & computer 2003*. Springer, 187–196.

[38] Behnam Hedayatnia, Seokhwan Kim, Yang Liu, Karthik Gopalakrishnan, Mihail Eric, and Dilek Hakkani-Tur. 2020. Policy-Driven Neural Response Generation for Knowledge-Grounded Dialogue Systems. *arXiv preprint arXiv:2005.12529* (2020).

[39] Samuel Holmes, Anne Moorhead, Raymond Bond, Huiru Zheng, Vivien Coates, and Michael McTear. 2019. Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces?. In *Proceedings of the 31st European Conference on Cognitive Ergonomics*. 207–214.

[40] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*.

[41] Chin-Yuan Huang, Ming-Chin Yang, Chin-Yu Huang, Yu-Jui Chen, Meng-Lin Wu, and Kai-Wen Chen. 2018. A Chatbot-supported smart wireless interactive healthcare system for weight control and health promotion. In *2018 IEEE international conference on industrial engineering and engineering management (IEEM)*. IEEE, 1791–1795.

[42] Becky Inkster, Shubhankar Sarda, and Vinod Subramanian. 2018. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth* 6, 11 (2018), e12106.

[43] Raquel Justo, Leila Ben Letaifa, Javier Mikel Olaso, Asier López-Zorrilla, Mikel Develasco, Alain Vázquez, and M Inés Torres. 2021. A Spanish Corpus for Talking to the Elderly. In *Conversational Dialogue Systems for the Next Decade*. Springer, 183–192.

[44] Raquel Justo, Leila Ben Letaifa, Cristina Palmero, Eduardo Gonzalez-Fraile, Anna Torp Johansen, Alain Vázquez, Stephan Schlögl Gennaro Cordasco, Begoña Fernández-Ruanova, Micaela Silva, Sergio Escalera, Mikel deVelasco, Joffre Tenorio-Laranga, Anna Esposito, Maria Korsnes, and M. Inés Torres. 2020. Analysis of the interaction between elderly people and a simulated virtual coach. *Journal of Ambient Intelligence and Humanized Computing* 11 (2020), 6125–6140.

[45] Dipesh Kadariya, Revathy Venkataramanan, Hong Yung Yip, Maninder Kalra, Krishnaprasad Thirunarayanan, and Amit Sheth. 2019. kBot: Knowledge-Enabled Personalized Chatbot for Asthma Self-Management. In *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*. IEEE, 138–143.

[46] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *Computing Research Repository* arXiv:1412.6980 (2014). https://arxiv.org/abs/1412.6980

[47] Rafal Kocielnik, Lillian Xiao, Daniel Avrahami, and Gary Hsieh. 2018. Reflection companion: A conversational system for engaging users in reflection on physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–26.

[48] Mark A Kramer. 1991. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal* 37, 2 (1991), 233–243.

[49] Ilya Kulikov, Alexander H Miller, Kyunghyun Cho, and Jason Weston. 2018. Importance of a search strategy in neural dialogue modelling. *Computing Research Repository* arXiv:1811.00907 (2018). https://arxiv.org/abs/1811.00907

[50] Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie Y S Lau, and Enrico Coiera. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association* 5, 9 (2018), 1248–1258. https://doi.org/10.1093/jamia/ocy072

[51] Lei Le, Andrew Patterson, and Martha White. 2018. Supervised autoencoders: Improving generalization performance with unsupervised regularizers. *Advances in Neural Information Processing Systems* 31 (2018), 107–117.

[52] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *Computing Research Repository* arXiv:1510.03055 (2015). http://arxiv.org/abs/1510.03055

[53] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A Persona-Based Neural Conversation Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 994–1003.

[54] Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial Learning for Neural Dialogue Generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2157–2169.

[55] Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087* (2019).

[56] Qian Lin, Souvik Kundu, and Hwee Tou Ng. 2020. A Co-Attentive Cross-Lingual Neural Model for Dialogue Breakdown Detection. In *Proceedings of the 28th International Conference on Computational Linguistics*. 4201–4210.

[57] Zhaojiang Lin, Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Yejin Bang, Etsuko Ishii, and Pascale Fung. 2020. XPersona: Evaluating Multilingual Personalized Chatbot. *Computing Research Repository* arXiv:2003.07568 (2020). https://arxiv.org/abs/2003.07568

[58] Pierre Lison, Jörg Tiedemann, Milen Kouylekov, et al. 2019. Open subtitles 2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *LREC 2018, Eleventh International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA).

[59] Bing Liu and Ian Lane. 2018. End-to-end learning of task-oriented dialogs. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. 67–73.

[60] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *EMNLP*.

[61] Yang Liu, Kun Han, Zhao Tan, and Yun Lei. 2017. Using context information for dialog act classification in dnn framework. In *Proceedings of the 2017 conference on empirical methods in natural language processing*. 2170–2178.

[62] Asier López Zorrilla, Mikel deVelasco Vázquez, and Raquel Justo. 2020. Euskarazko elkarrizketa sistema automatikoa sare neuronalen bidez (A neural dialogue system in Basque). *Ekaia. EHUko Zientzia eta Teknologia aldizkaria* (2020).

[63] Asier López Zorrilla, Mikel deVelasco Vázquez, and M. Inés Torres. 2021. A Differentiable Generative Adversarial Network for Open Domain Dialogue. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*. Springer Singapore, Singapore, 277–289. https://doi.org/10.1007/978-981-15-9323-9_24

[64] Asier López Zorrilla, Mikel de Velasco Vázquez, Jon Irastorza, Javier Mikel Olaso Fernández, Raquel Justo Blanco, and María Inés Torres Barañano. 2018. Empathic: Empathic, expressive, advanced virtual coach to improve independent healthy-life-years of the elderly. (2018).

[65] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *Computing Research Repository* arXiv:1711.05101 (2017). https://arxiv.org/abs/1711.05101

[66] Xueming Luo, Marco Shaojun Qin, Zheng Fang, and Zhe Qu. 2021. Artificial Intelligence Coaches for Sales Agents: Caveats and Solutions. *Journal of Marketing* 85, 2 (2021), 14–32.

[67] James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1. Oakland, CA, USA, 281–297.

[68] Carol Ann Maher, Courtney Rose Davis, Rachel Grace Curtis, Camille Elizabeth Short, and Karen Joy Murphy. 2020. A Physical Activity and Diet Program Delivered by Artificially Intelligent Virtual Health Coach: Proof-of-Concept Study. *JMIR mHealth and uHealth* 8, 7 (2020), e17558.

[69] Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised Evaluation of Interactive Dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 225–235.

[70] Shiwali Mohan, Anusha Venkatakrishnan, and Andrea L. Hartzler. 2020. Designing an AI Health Coach and Studying Its Utility in Promoting Regular Aerobic Exercise. *ACM Trans. Interact. Intell. Syst.* 10, 2, Article 14 (May 2020), 30 pages. https://doi.org/10.1145/3366501

[71] César Montenegro, Asier López Zorrilla, Javier Mikel Olaso, Roberto Santana, Raquel Justo, Jose A Lozano, and María Inés Torres. 2019. A dialogue-act taxonomy for a virtual coach designed to improve the life of elderly. *Multimodal Technologies and Interaction* 3, 3 (2019), 52.

[72] César Montenegro, Roberto Santana, and José Antonio Lozano. 2019. Data generation approaches for topic classification in multilingual spoken dialog systems. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. 211–217.

[73] Rostislav Nedelchev, Jens Lehmann, and Ricardo Usbeck. 2020. Language Model Transformers as Evaluators for Open-domain Dialogues. In *Proceedings of the 28th International Conference on Computational Linguistics*. 6797–6808.

[74] Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics* 6 (2018), 373–389.

[75] Stefan Olafsson, Teresa K. O'Leary, and Timothy W. Bickmore. 2020. Motivating Health Behavior Change with Humorous Virtual Agents. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents* (Virtual Event, Scotland, UK) *(IVA '20)*. Association for Computing Machinery, New York, NY, USA, Article 42, 8 pages. https://doi.org/10.1145/3383652.3423915

[76] Daniel Ortega, Chia-Yu Li, Gisela Vallejo, Pavel Denisov, and Ngoc Thang Vu. 2019. Context-aware neural-based dialog act classification on automatically generated transcriptions. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7265–7269.

[77] Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 1703–1714. https://www.aclweb.org/anthology/2020.acl-main.156

[78] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[79] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

[80] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019), 9.

[81] K. Ralston, Y. Chen, H. Isah, and F. Zulkernine. 2019. A Voice Interactive Multilingual Student Support System using IBM Watson. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. 1924–1929.

[82] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8689–8696.

[83] Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/2004.09813

[84] Deborah Richards and Patrina Caldwell. 2017. Improving health outcomes sooner rather than later via an interactive website and virtual specialist. *IEEE journal of biomedical and health informatics* 22, 5 (2017), 1699–1706.

[85] Mario Rodríguez-Cantelar, Luis Fernando D'Haro, and Fernando Matía. 2020. Automatic Evaluation of Non-task Oriented Dialog Systems by Using Sentence Embeddings Projections and Their Dynamics. In *Conversational Dialogue*

Systems for the Next Decade. Springer, 71–84.

[86] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *Computing Research Repository* arXiv:2004.13637 (2020). https://arxiv.org/abs/2004.13637

[87] Abdelrhman Saleh, Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, and Rosalind Picard. 2019. Hierarchical reinforcement learning for open-domain dialog. *Computing Research Repository* arXiv:1909.07547 (2019). https://arxiv.org/abs/1909.07547

[88] Motoki Sano, Hiroki Ouchi, and Yuta Tsuboi. 2018. Addressee and Response Selection for Multilingual Conversation. In *Proceedings of the 27th International Conference on Computational Linguistics*. 3631–3644.

[89] Amit Saxena, Mukesh Prasad, Akshansh Gupta, Neha Bharill, Om Prakash Patel, Aruna Tiwari, Meng Joo Er, Weiping Ding, and Chin-Teng Lin. 2017. A review of clustering techniques and developments. *Neurocomputing* 267 (2017), 664–681.

[90] Stephan Schlögl, Gavin Doherty, Nikiforos Karamanis, and Saturnino Luz. 2010. WebWOZ: a wizard of oz prototyping framework. In *Proceedings of the 2nd ACM SIGCHI symposium on Engineering interactive computing systems*. 109–114.

[91] Helmut Schröder, Montserrat Fitó, Ramón Estruch, Miguel A. Martínez-González, Dolores Corella, Jordi Salas-Salvadó, Rosa Lamuela-Raventós, Emilio Ros, Itziar Salaverría, Miquel Fiol, José Lapetra, Ernest Vinyoles, Enrique Gómez-Gracia, Carlos Lahoz, Lluis Serra-Majem, Xavier Pintó, Valentina Ruiz-Gutierrez, and María-Isabel Covas. 2011. A Short Screener Is Valid for Assessing Mediterranean Diet Adherence among Older Spanish Men and Women. *The Journal of Nutrition* 141, 6 (04 2011), 1140–1145. https://doi.org/10.3945/jn.110.135566 arXiv:https://academic.oup.com/jn/article-pdf/141/6/1140/31113898/1140.pdf

[92] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1715–1725.

[93] Manex Serras, María Inés Torres, and Arantza Del Pozo. 2019. User-aware dialogue management policies over attributed bi-automata. *Pattern Analysis and Applications* 22, 4 (2019), 1319–1330.

[94] Michael Steinbach, Levent Ertöz, and Vipin Kumar. 2004. The challenges of clustering high dimensional data. In *New directions in statistical physics*. Springer, 273–309.

[95] Joffre Tenorio-Laranga, Begoña Fernández-Ruanova, M. Inés Torres, Raquel Justo, Alfredo Alday, and Josu Llano Hernaiz. 2019. Designing a virtual coach: Involvement of end-users from early design to prototype. *Journal of Ambient Intelligence and Humanized Computing* 19, 4 (2019), 207. https://doi.org/10.5334/ijic.s3207

[96] Natalia Tomashenko, Christian Raymond, Antoine Caubrière, Renato De Mori, and Yannick Estève. 2020. Dialogue history integration into end-to-end signal-to-concept spoken language understanding systems. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8509–8513.

[97] María Inés Torres, Javier Mikel Olaso, César Montenegro, Roberto Santana, A Vázquez, Raquel Justo, José Antonio Lozano, Stephan Schlögl, Gérard Chollet, Nazim Dugan, et al. 2019. The empathic project: mid-term achievements. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. 629–638.

[98] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[99] Alain Vázquez. 2019. *EMPATHIC-NLG: un generador de lenguaje natural adaptado al coaching (EMPATHIC-NLG: a natural language generator adapted to coaching)*. Master's thesis. University of the Basque Country UPV/EHU.

[100] Jian Wang, Junhao Liu, Wei Bi, Xiaojiang Liu, Kejing He, Ruifeng Xu, and Min Yang. 2020. Dual Dynamic Memory Network for End-to-End Multi-turn Task-oriented Dialog Systems. In *Proceedings of the 28th International Conference on Computational Linguistics*. 4100–4110.

[101] John Whitmore. 1992. Growing human potential and purpose: The principles and practice of coaching and leadership.

[102] Jason D Williams, Kavosh Asadi Atui, and Geoffrey Zweig. 2017. Hybrid Code Networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 665–677.

[103] Jason D Williams and Steve Young. 2007. Partially observable Markov decision processes for spoken dialog systems. *Computer Speech & Language* 21, 2 (2007), 393–422.

[104] Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *Computing Research Repository* arXiv:1901.08149 (2019). https://arxiv.org/abs/1901.08149

[105] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A New Chatbot for Customer Service on Social Media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3506–3510. https://doi.org/10.1145/3025453.3025496

[106] Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2019. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307* (2019).

[107] Jingwen Zhang, Yoo Jung Oh, Patrick Lange, Zhou Yu, and Yoshimi Fukuoka. 2020. Artificial Intelligence Chatbot Behavior Change Model for Designing Artificial Intelligence Chatbots to Promote Physical Activity and a Healthy Diet. *Journal of medical Internet research* 22, 9 (2020), e22845.

[108] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too?. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2204–2213.

[109] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

[110] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. 1996. BIRCH: An Efficient Data Clustering Method for Very Large Databases. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data* (Montreal, Quebec, Canada) *(SIGMOD '96)*. Association for Computing Machinery, New York, NY, USA, 103–114. https://doi.org/10.1145/233269.233324

[111] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

A CLUSTER SEQUENCE MODELING RESULTS

Table 13 shows a the results of the cluster modeling task. The accuracy and top N accuracy (with N=3) on the test set are shown for each language. The GRU and three N gram models are compared.

Table 13. Accuracy and top N accuracy (with N=3) obtained by the cluster sequence modeling models across the four languages on the test set.

	English		Spanish		French		Norwegian	
	Acc.	Top N acc.	Acc.	Top N acc.	Acc.	Top N acc.	Acc.	Top N acc.
GRU	0.350	0.581	0.346	0.567	0.327	0.592	0.356	0.616
Bigrams	0.243	0.479	0.247	0.475	0.248	0.535	0.231	0.522
Trigrams	0.183	0.352	0.188	0.376	0.177	0.413	0.173	0.396
4-grams	0.147	0.304	0.155	0.299	0.151	0.349	0.146	0.323

B THE INFLUENCE OF THE NUMBER OF CLUSTERS

The number of clusters was set to 60 in the K-Means algorithm. Let us show the relation of the selected number of clusters and the number of turns per cluster, the WDH system's accuracy at the next utterance classification task and the F1 score at dialogue act classification from the cluster index. In order to analyse how balanced the number of turns per cluster is, we computed its coefficient of variation (the ratio of the standard deviation to the mean). A lower coefficient of variation implies that the clusters are more balanced, i.e., that they include a more similar number of turns each; whereas larger values indicate that some clusters are very populated while others contain very few turns inside. Lower values are therefore preferred, since they should allow a better modeling of the cluster flow, as this data set would be more balanced. More information on how the next utterance classification accuracy and the F1 score of dialogue act classification from clusters are computed can be found at Sections 7.4 and 9.2, respectively. Figure 14 shows the influence of the number of clusters for these metrics. For simplicity, we only show the results of these experiment in English, but they follow a similar tendency in all the versions of the corpus. The metrics were obtained on the test partition.

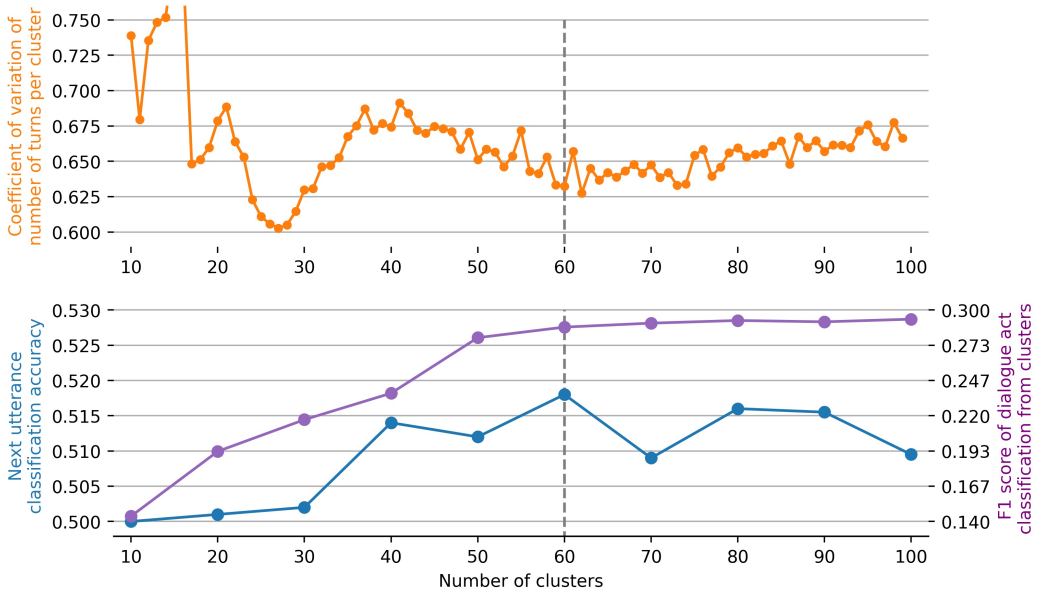


Fig. 14. Different metrics in terms of the selected number of clusters. On top, the coefficient of variation of the number of turns per cluster. At the bottom, the next utterance classification accuracy in blue (the scale is on the left), and the F1 score of dialogue act classification from clusters in purple (the scale is on the right).

Regarding the coefficient of variation of the number of turns per cluster, there seem to be a first and very notorious local minimum at around 25-30 clusters. Then the values go up at 40 clusters and they are reduced again, even though slightly at around 60-70 clusters. We decided to select 60 instead of 25 or 30 due to the behaviour of the other metrics. The F1 score of dialogue act classification from clusters gets better as the number of clusters increase. However, this metric stabilizes at around 50-60 clusters, and it does not improve significantly from there on. Last, the next utterance classification accuracy is the noisiest metric. Anyway, it can be seen that it is worse with less clusters (10-30), and then it improves after 40, with the maximum at 60. Thus we believe that the choice of 60 clusters represents a good balance between all these metrics.

C QUESTIONNAIRES

Tables 14 and 15 show the Chatbot Usability Questionnaire and the Hedonic Feelings Questionnaire, respectively.

Table 14. Chatbot Usability Questionnaire.

Question code	Question
CUQ-1	The chatbot's personality was realistic and engaging.
CUQ-2	The chatbot seemed too robotic.
CUQ-3	The chatbot was welcoming during initial setup.
CUQ-4	The chatbot seemed very unfriendly.
CUQ-5	The chatbot explained its scope and purpose well.
CUQ-6	The chatbot gave no indication as to its purpose.
CUQ-7	The chatbot was easy to navigate.
CUQ-8	It would be easy to get confused when using the chatbot.
CUQ-9	The chatbot understood me well.
CUQ-10	The chatbot failed to recognise a lot of my inputs.
CUQ-11	Chatbot responses were useful, appropriate and informative.
CUQ-12	Chatbot responses were not relevant.
CUQ-13	The chatbot coped well with any errors or mistakes.
CUQ-14	The chatbot seemed unable to handle any errors.
CUQ-15	The chatbot was very easy to use.
CUQ-16	The chatbot was very complex.

Table 15. Hedonic Feelings Questionnaire.

Question code	Question
HFQ-1	I think the communication with the agent was extraordinary.
HFQ-2	I think the communication with the agent was boring.
HFQ-3	I think the communication with the agent was innovative.
HFQ-4	I think the communication with the agent was disappointing.
HFQ-5	I think the communication with the agent was thrilling.
HFQ-6	I think the communication with the agent was trivial.
HFQ-7	I think the communication with the agent was stimulant.
HFQ-8	I think the communication with the agent was depressing.
HFQ-9	I think the communication with the agent was reassuring.
HFQ-10	I think the communication with the agent was stressful.

D THE SET OF DIALOGUE ACTS

The dialogue acts employed to perform the dialogue act classification in Section 9.2 and the analysis of the dialogue flow in Section 9.3 are shown in Table 16.

Table 16. Dialogue acts and their abbreviations.

Dialogue act	Description
Hello	Hello, salutation.
Ask name	Ask about the user's name and spelling.
Patience request	Patience request.
Self-intro	The system presents itself.
Know coaching?	Ask about the user's knowledge about coaching.
Echo	Repeat something said by the user, to transmit empathy and understanding.
Open Q	An open question about the user.
Yes/no Q	A yes/no question about the user.
Music	A question/statement about music.
Travel	A question/statement about travelling.
Other hobbies	A question/statement about a hobby that is not music nor travelling.
I understand	Explicitly tell the user that their message has been understood.
Clarify	Ask for a clarification.
Neg feedback	Disagree with the user or show a negative/non-positive opinion.
Pos feedback	Show a positive opinion.
Agreement	Agree with the user.
Topic	Open, close or choose a new topic.
Current situation	Questions about the current situation of the user regarding their goal.
GSQ-IS	Goal Setting Question - Ideal Situation. Ask the user which would be the ideal situation in connection with their goal
GSQ-Obj	Goal Setting Question - Objective. Ask the user to define their goal.
MQ	Motivational Question.
ORQ	Obstacles/Resources Question. Questions to find out which obstacles that hinder the achievement of the goal, and the possible resources to overcome them.
PAQ	Plan Action Question. Question to define a plan that brings the user closer to their objective.
Thanking	Thank the user.
Farewell	Say goodbye.
Other	A system turn that was not classifiable as any of the aforementioned dialogue acts.