Toward Data-Driven Collaborative Dialogue Systems: The JILDA Dataset

Irene Sucameli*
Università di Pisa

Alessandro Lenci**
Università di Pisa

Bernardo Magnini†
Fondazione Bruno Kessler

Manuela Speranza‡
Fondazione Bruno Kessler

Maria Simi§
Università di Pisa

Today's goal-oriented dialogue systems are designed to operate in restricted domains and with the implicit assumption that the user goals fit the domain ontology of the system. Under these assumptions dialogues exhibit only limited collaborative phenomena. However, this is not necessarily true in more complex scenarios, where user and system need to collaborate to align their knowledge of the domain in order to improve the conversation and achieve their goals.

To foster research on data-driven collaborative dialogues, in this paper we present JILDA, a fully annotated dataset of chat-based, mixed-initiative Italian dialogues related to the job-offer domain. As far as we know, JILDA is the first dialogic corpus completely annotated in this domain. The analysis realised on top of the semantic annotations clearly shows the naturalness and greater complexity of JILDA's dialogues. In fact, the new dataset offers a large number of examples of pragmatic phenomena, such as proactivity *(i.e., providing information not explicitly requested) and* grounding*, which are rarely investigated in AI conversational agents based on neural architectures. In conclusion, the annotated JILDA corpus, given its innovative characteristics, represents a new challenge for conversational agents and an important resource for tackling more complex scenarios, thus advancing the state of the art in this field.*

1. Introduction

In recent years, mostly driven by the high performance achieved by deep learning approaches in Natural Language Processing, there has been a resurgence of interest for systems that are able to assist people in a number of tasks, interacting in a natural way. However, reproducing the peculiarity and complexity of human-human dialogues

* Department of Computer Science - Largo Bruno Pontecorvo 3, 56127 Pisa, Italy. E-mail: irene.sucameli@phd.unipi.it
** Department of Philology, Literature and Linguistics - Via Santa Maria, 56126 Pisa, Italy. E-mail: alessandro.lenci@unipi.it
† Natural Language Processing Group - Via Sommarive 18 Povo, 38123 Trento, Italy. E-mail: magnini@fbk.eu
‡ Natural Language Processing Group - Via Sommarive 18 Povo, 38123 Trento, Italy. E-mail: manspera@fbk.eu
§ Department of Computer Science - Largo Bruno Pontecorvo 3, 56127 Pisa, Italy. E-mail: simi@di.unipi.it

poses a number of scientific challenges to current conversational AI approaches, and, more generally, to computational linguistics. In this paper we present JILDA, a corpus of human-human dialogues collected with the purpose of investigating linguistic variability and collaborative phenomena in *goal-oriented dialogues*, which imply a collaborative effort to plan actions among the interlocutors in order to achieve a certain communicative goal.

9. Applicant: *Nel frattempo potrei specificarti le mie preferenze a livello geografico? Potrebbero aiutarti nel targetizzarmi meglio*

10. Navigator: *SÃň, perfetto! Grazie*

11. Applicant: *Attualmente vivo in Toscana: sono disponibile a trasferirmi in altre regioni ma anche allâĂŹestero non ho problemi di mobilitÃă o limiti da questo punto di vista*

12. Navigator: *Potrei avere due offerte che mi piacerebbe proporti. Entrambe riguardano tirocini post-laurea, uno come assistente capocommessa in una azienda edile a Pistoia, e l'altra come allievo direttore a Milano presso Compass.*

(...)

15. Applicant: *Non riesco a capire bene che cosa significhi "allievo direttore"*

16. Navigator: *Certo! Le principali mansioni legate a questo impiego riguardano la pianificazione del budget e del conto economico dell'azienda. Il settore Ãĺ quello alimentare quindi si tratta di compilare ordini e derrate alimentari, oltre che garantire la sicurezza sul lavoro e quella alimentare.*

17. Navigator: *Compiti gestionali sarebbero sicuramente al centro del lavoro.*

18. Navigator: *Ti sembra piÃź chiaro? Posso dirti altro?*

19. Applicant: *Capisco. Mi sembra interessante*

(...)

21. Navigator: *Trattandosi di un tirocinio post-laurea direi che la formazione sarÃă una componente importante.*

22. Applicant: *Capisco. CâĂŹÃĺ una deadline per fare domanda?*

(...)

28. Applicant: *Capisco. Potresti darmi il contatto dellâĂŹazienda? In modo tale da approfondire e mettermi in contatto diretto con loro*

9. Applicant: *In the meantime, should I specify my geographic preferences? They could help you target me better*

10. Navigator: *Yes, perfect! Thank you*

11. Applicant: *At the moment I live in Tuscany: I'm available to move to other regions and even abroad I don't have mobility problems or limitations from this point of view*

12. Navigator: *I may have two offers that I would like to propose to you. Both involve post-graduate internships, one as an assistant prime contractor in a construction company in Pistoia, and the other as a junior director in Milan at Compass.*

(...)

15. Applicant: *I can't quite understand what "junior director" means*

16. Navigator: *Sure! The main tasks related to this job concern the planning of the budget and the income statement of the company. The area is the food sector so it's a question of filling orders and foodstuffs, as well as guaranteeing work and food safety.*

17. Navigator: *Management tasks would certainly be the core of the work.*

18. Navigator: *Is it more clear now? Can I tell you more?*

19. Applicant: *I see. It seems interesting*

(...)

21. Navigator: *Since this is a post-graduate internship I would say that training will be an important component.*

22. Applicant: *I see. Is there an application deadline?*

(...)

28. Applicant: *I see. Could you give me the company's contact? This way I can take a closer look and contact them directly*

Goal-oriented dialogues contain interactions governed by shared conventions (see, for instance the work of (Grice 1975) on conversational maximes), which involve knowledge about the *pragmatics* of language (Levinson 1983), i.e., the context in which they

are produced and the speakers' communicative intentions. In this paper we focus on two pragmatic phenomena that are relevant in goal-oriented dialogues: *proactivity* and *grounding*. To give an intuition of what proactivity and grounding are, and how they are pervasive in human dialogues, let's consider the following extract, from a goal-oriented dialogue from the JILDA corpus (full version available in Appendix), where a navigator and an applicant have to find a satisfactory match between a set of job offers and the applicant's CV.

Proactivity (Balaraman and Magnini 2020b) occurs when an interlocutor offers information which was not explicitly requested, with the intention of facilitating the achievement of the conversational goal. As an example, at lines 9 and 11 of the dialogue, the applicant offers information which was not asked by the navigator (i.e., its geographical working preferences), but is assumed to facilitate the search of an appropriate job offer. The navigator, too, at line 16 provides details about a company which were actually not required by the applicant question at line 15. Even in this case the purpose is facilitating the match of a job offer with the applicant's requirements.

Grounding (Clark and Schaefer 1987; Clark and Brennan 1991; Hough and Schlangen 2017) is the process through which participants in a dialogue build and keep themselves aligned to a common knowledge ground, formed by interlocutors' shared information. Depending on the state of the dialogue, it is possible to identify several types of grounding (Traum 1999; Hough et al. 2015), such as, for instance, *feedback* and *repair*, which allow participants to demonstrate their understanding of the conversation or to correct potential misunderstandings.

Grounding is particularly relevant in goal-oriented dialogue (Mushin et al. 2003), where the participants are not supposed to share part of their knowledge. In our example dialogue from the JILDA corpus, grounding occurs in several forms. At line 15 it is the applicant who poses a clarification question *I canâĂŹt quite understand what "junior director" means*. At line 18 the navigator asks for confirmation *s it more clear now? Can I tell you more?*, while at lines 19, 22 and 28 the applicant explicitly recognises to be aligned with the navigator.

Although grounding and proactivity are pervasive in human-human dialogue, both are largely under represented in current data-driven, goal-oriented, dialogue systems. This is related to the fact that both phenomena are scarcely present in training data, which, in turn, may depend on the design choices adopted by developers for the collection of dialogues. Two design choices seem to be relevant: (i) some acquisition methodologies (e.g., Wizard of Oz) constrain participants in the data collection to follow pre-defined dialogue scripts, resulting in dialogues that are quite repetitive and poor in natural pragmatic phenomena; (ii) in most cases the domain of conversation is oversimplified with respect to the real world (e.g., when booking restaurants, they are described with few characteristics), resulting in a reduced need for grounding between the system and the user.

JILDA consists of goal-oriented, chat based, Italian dialogues related to the job-offer domain. The corpus is fully annotated with semantic information, such as dialogue acts and entities, as well as proactive phenomena. It is important to underline that the annotation of proactivity has been included in the dataset to better capture the complexity of a natural, human-human dialogue. This annotation therefore represents an important characteristic of the dataset itself and is useful for conducting a linguistic analysis of the Italian language, but it is not designed to develop a system capable of producing proactive behaviour.

We describe in detail the annotation methodology adopted in JILDA and analyse and discuss the major novelties introduced in the corpus, showing high presence of

pragmatic phenomena, including grounding and proactivity. We expect that JILDA can be used to train neural dialogue models for the Italian language (JILDA is a quite new resource for this language), thereby pushing the scientific community toward more natural and effective conversational systems.

2. Background on Goal-oriented Dialogue

In this section we introduce relevant background on goal-oriented dialogues, which may help to appreciate the novelty of the JILDA corpus. First we highlight some of the characteristics of goal-oriented dialogues, then we briefly introduce some notion relevant to the realisation of automatic goal-oriented dialogue systems, and, finally, we focus on the presence of collaborative behaviours in some datasets developed to train conversational systems.

2.1 Human-human Goal-oriented Dialogue

The purpose of a typical task-oriented dialogue is to retrieve pieces of information that are supposed to correspond to user needs (e.g., booking a restaurant, finding how to open a bank account, check the weather tomorrow, etc.). It is usually assumed that the user has a rather clear goal in mind, which is then elicited by an operator during the dialogue. The operator in fact may ask questions to the user attempting to reduce the search space and to focus on those objects that fit the user goals. On the other side, the user may also intervene in the dialogue to clarify and refine the goals of the conversation. Once objects that satisfy the user needs are retrieved, an action can be executed, such as booking a restaurant, or blocking a credit card. A goal-oriented dialogue may terminate either when the goal has been achieved (e.g., a reservation has been confirmed), or when the goal can not be achieved, because it was not possible to find a match with the user needs.

As an example of human-human goal-oriented dialogue, let's consider the following excerpt from Nespole (Mana et al. 2004, 2003), a corpus consisting of spoken interactions between a professional agent and a client about vacation planning in the Trentino region.

1. Client: *Good morning; could you suggest any village in the Val di Fiemme to me; where itâĂŹs possible to skate for example; that is does any skating rink exist in the Val di Fiemme;*

2. Agent: *yes; in the whole of Val di Fiemme there are some outdoor skating rinks; where you can skate usually in the afternoon; in some rinks even in the morning; and then right in Cavalese thereâĂŹs a skating rink an ice rink; where even some courses are organized; where they also hold hockey or skating shows; and itâĂŹs indoors.*

What is interesting for our purposes is the collaborative attitude of both the Client and the Agent. Particularly, the travel agent proactively provides indications both about the opening time of skating rinks and about skating courses, which were not explicitly requested by the customer. Proactivity is a peculiar characteristics of human-human dialogues, through which the Agent anticipates the expected requests of the user, this way facilitating the achievements of the dialogue goals.

2.2 Goal-oriented Dialogue Systems

Task-oriented dialogue systems aim to assist users to accomplish a task (e.g., booking a flight, making a restaurant reservation and playing a song) through dialogue in natural language, either in a spoken or written form. As in most current approaches, we assume a system involving a pipeline of components - see Figure 1, from (Deriu et al. 2021) - where the user utterance is first processed by an Automatic Speech Recognition (ASR) module and then processed by a Natural Language Understanding (NLU) component, which interprets the user's needs (Louvan and Magnini 2020). Then a Dialogue State Tracker (DST) (Balaraman, Sheikhalishahi, and Magnini 2021) accumulates the dialogue information as the conversation progresses and may query a domain knowledge base to obtain relevant data. A dialogue policy manager then decides the next action to be executed and, finally, a Natural Language Generation (NLG) component produces the actual response to the user.

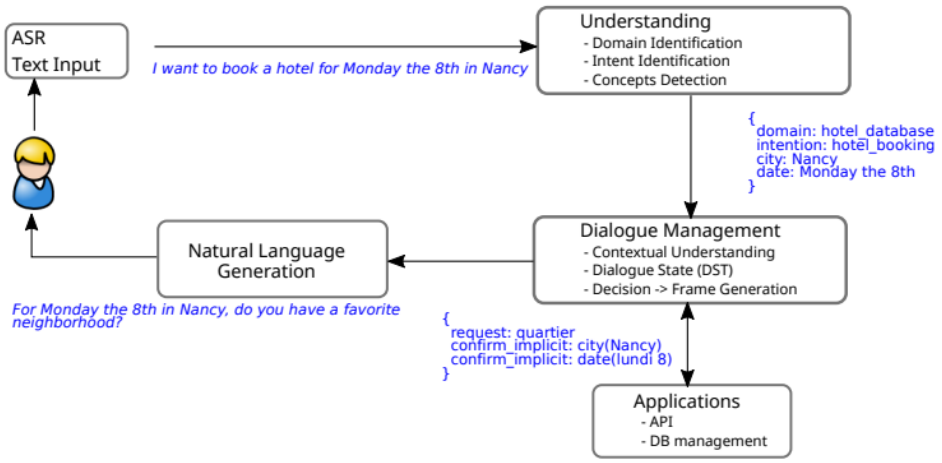

Figure 1
A standard architecture of a task-oriented dialogue system

In order to reproduce collaborative behaviours, the most relevant component is the dialogue manager, which has to decide whether a collaborative action is appropriate for the current dialogue turn, given the dialogue history and the user beliefs (i.e., the supposed user goals). For a dialogue manager the question is how to learn proactive behaviours, including knowledge about turns in which the system should be proactive, and when it should not, how to determine the information that should be proactively offered to the user, and the appropriate amount of such information (e.g., offering too much information may result in a excessive cognitive effort for the user). Similar questions apply to grounding, where the dialogue manager has to constantly monitor the level of grounding with the user, and, in case this is not satisfactory, has to take the initiative to restore it to an optimal level.

Given the inherent complexity of collaborative behaviours, it is not surprising that current dialogue systems still have limited capacities in this respect. The issue of reproducing collaborative behaviours is even more evident for a data-driven dialogue state tracker, which is assumed to learn dialogue behaviours from annotated dialogues.

In this case, the availability of dialogues displaying reach enough linguistic phenomena is crucial.

2.3 Datasets for Goal-oriented Dialogue

As dialogic annotated corpora are at the core of the capacity to learn dialogue models, this section introduces the most important available datasets, focusing on the presence of collaborative phenomena. As a case study, we have selected *WoZ* and *MultiWoZ*, two datasets developed in recent years, which are considered as benchmarks for developing deep learning methods for dialogue state tracking.

WoZ is a popular dataset for restaurant booking in Cambridge, collected using the Wizard of Oz approach, where the user and the wizard contribute a single turn to each dialogue (Wen et al. 2017). (Mrkšić et al. 2017) expanded WoZ into WoZ2.0, consisting of 1,200 dialogues. Then, MultiWOZ2.1 (Budzianowski et al. 2018) further extends WoZ including dialogues in multiple domains. To this aim, the dataset developers explicitly encouraged goal changes, in order to model more realistic conversations. Different versions of MultiWOZ2.1 have been recently published, addressing annotation errors occurring in the original dataset (Ramadan, Budzianowski, and Gasic 2018; Budzianowski et al. 2018; Eric et al. 2020; Zang et al. 2020). MultiWoZ2.1 contains 10,438 dialogues, covering several different domains (e.g., restaurants, hotels, trains and attractions).

Both datasets have been collected through the Wizard of Oz approach, (Kelley 1984), where a human (the "wizard") plays the role of the computer within a simulated human-computer conversation, and, crucially the other speakers are not aware to talk to a human. The following is an example of a dialogue script provided to the "user" in the Wizard of Oz collection setting.

1. User: *You are looking for a <place to stay>. The hotel should be in the <cheap> price range.*

2. User: *The hotel should <include free parking> and should <include free wifi>*

3. User: *Once you find the <hotel> you want to book it for <6 people> and <3 nights> starting from <tuesday>*

4. User: *If the booking fails how about <2 nights>*

5. User: *Make sure you get the <reference number>*

The dialogue script is typically filled in using placeholders in a template (shown in *<italics>* in the example). It is worth to notice the amount of details that are present in the dialogue description, details that could influence the production of the user utterance for a given turn, and induce to follow a structure similar to that of the dialogue script. After being collected through Wizard of Oz, turns of each dialogue are annotated with the corresponding *dialogue state*, consisting of an intent and a set of slot-value pairs. The following is an example of the annotation provided in a portion of a MultiWoZ 2.0 dialogue:

1. User: *I would like a moderately priced restaurant in the west part of town.*
INFORM(PRICE=MODERATE, AREA WEST)

2. System: *here are three moderately priced restaurants in the west part of town. Do you prefer Indian Italian or British?*
REQUEST(FOOD)

3. User: *Can I have the address and phone number of the Italian location?*
INFORM(PRICE=MODERATE, AREA=WEST, FOOD=ITALIAN)
REQUEST(ADDRESS, PHONE-NUMBER)

Neither proactivity nor grounding are annotated in WoZ and MultiWoz. A recent study (Balaraman and Magnini 2020a), estimated that the amount of the system proactive behaviours in MultiWoz is rather low. In fact, out of 143,048 dialogue turns in the corpus, only 325 proactive turns were found with a clear proactive pattern. Although this might be an underestimation (as proactivity is not annotated in MultiWoz and it is not trivial to search for it), this is much less than we can reasonably expect in human-human goal-oriented dialogues, as the example reported in the introduction shows. Being poorly represented in the corpus, proactive behaviours can hardly be learnt by dialogue state tracking and dialogue policy models, motivating the need of richer dialogue annotations, such as those proposed in JILDA.

Other popular datasets used for dialogue state tracking include the schema-guided dataset (Shah et al. 2018), collected using a bootstrapping approach, and the TreeDST dataset (Cheng et al. 2020), with conversations covering 10 domains. These datasets mainly focus on the problem of managing a conversational domain with scarcity of training data (e.g., the problem of managing unseen slot values), proposing architectures (e.g., zero shot learning) that are robust enough for such situations. To the best of our knowledge, there is no much attention to explore collaborative phenomena in dialogue.

Finally, it is worth briefly reporting about the performance that state-of-the-art models achieve on the dialogue state tracking task. MultiWoz is probably the dataset mostly used to train a dialogue state tracker model, and several deep learning architectures have been experimented in the last years (Henderson, Thomson, and Young 2014; Balaraman and Magnini 2021), including methods proposed at various editions of the DST challenge (Henderson, Thomson, and Williams 2014). Performance are typically reported according to the *joint goal accuracy* of the model, i.e, the capacity of the model to correctly predict all dialogue states (slot-value pairs) in each turn of the dialogue. Current DST models, for instance TRADE (Wu et al. 2019), DST-QA (Zhou and Small 2019) and CHAN-DST (Shan et al. 2020), achieve a performance in the order of 50% of joint goal accuracy.

The JILDA dataset, which will be described in detail in the next sections, builds on top of the experience accumulated by MultiWoz, proposing, however, a number of methodological improvements. First of all JILDA has been collected through Map-task, a methodology that allows the participants to express themselves with more naturalness (i.e., rich language variability) than in the Wizard of Oz setting, this way overcoming some of the limitations of current datasets. Second, the selected domain, job offers, is more complex than the MultiWoz domains, which should favour grounding phenomena among interlocutors. Finally, although we basically follow the MultiWoz annotation schema, we have added categories specifically tailored to mark dialogue collaborative phenomena.

3. JILDA

JILDA is a dataset of chat-based dialogues, produced by 50 Italian native speakers and related to the job-offer domain. The dataset, which is available on GitHub,[1] includes 525 mixed-initiative dialogues collected from human-human conversations in an experiment inspired by the Map-task methodology, where one participant played the role of job consultant (or "navigator") and the other the role of applicant, with the common goal of finding a good match between job offers and the applicantâĂŹs competences and expectations (Sucameli et al. 2020).

In a previous experiment we collected via Amazon Mechanical Turk another dataset of dialogues (Mturk), for the same domain and language as JILDA, using a template-based approach. Table 1 summarises the main characteristics of JILDA, highlighting the differences between this dataset with respect to the Mturk dataset.

Table 1
Comparison between MTurk's and JILDA's dialogues. Values marked with an asterisk are computed considering the average value of three JILDA's subsets, each including the same number of tokens as MTurk

	MTurk	**JILDA**
# dialogues	220	525
avg turns per dialogue	8	17
# tokens	45972	217132
# sentences	5201	20644
# utterances	3380	14509
# types	1975	6519
# lemmas	1605	4913
type/token ratio	0.043	0.072*
lemma/token ratio	0.035	0.056*
avg length sentences	9.24	10.52
avg length utterances	13.58	14.94

As shown by Table 1, JILDA is characterised by a great linguistic variability and lexical complexity that we tried to capture effectively during the subsequent annotation phase.

3.1 Annotation Guidelines

The JILDA annotation scheme relies on the MultiWOZ 2.1 one (Budzianowski et al. 2018). Differently from MultiWOZ however, we annotate both applicant and navigator utterances. In fact, one of the main characteristics of JILDA is to include mixed-initiative dialogues, where both participants involved in the conversation may ask and answer questions, or volunteer information, thus conveying useful data worth extracting. In the following we will use the most standard terms "system" and "user" to refer to navigator and applicant. In fact, JILDA was created with the idea of training a dialogic system on this domain. In this scenario, the system would cover the role of navigator, while

1 https://github.com/IreneSucameli/JILDA

the user would play the role of applicant. We annotate dialogue acts, which "represent the communicative intention behind a speakerâĂŹs utterance in a conversation" [Chakravarty, Chava, and Fox 2019], and slots, which are specific to the JILDA job-offer domain.

3.1.1 Dialogue Acts

For our annotation we considered six *Dialogue acts*, and we annotated both user's and system's utterances. Each act describes a specific communicative intention of the speaker. More specifically, the dialogue acts used for the annotation are:

- **greet**: the speaker expresses a greeting. Example:
  "Good morning, my name is Giulia and today I will be your navigator".
- **inform-basic**: the speaker provides information following a specific request. Example:
  sys: *"Tell me something about you: what type of studies have you done??"*
  usr: *"I graduated from classical high school and then got a degree in nursing"*
- **inform-proactive**: the speaker provides information that was not explicitly requested. For example, in the case below the system provides a piece of information (the email address) even if these data were not requested by the user:
  "Could you tell me where the company is located??"
  sys: *"The company is in Milan. You can get in touch with them with the email address info@azienda.com"*
- **request**: the speaker requests information:
  sys: *"Which sector would you like to work in?"*
- **select**: a) the system selects the job offer suitable for the user's profile or b) the user accepts the job offer. Example:
  sys: *"Ok I found an offer that meets your interests: it is a post-graduate internship in the food sector."*
- **deny**: the speaker is unable to satisfy a request. It includes, but is not limited to, categorizing cases in which the system does not find a suitable job offer for the user or the user does not accept the proposed offer. Example:
  usr: *"I don't think this offer works for me."*

Each sentence can be annotated with more than one dialogue act. For example, if the speaker, in addition to directly answering the interlocutor's question, volunteers additional information, the sentence is annotated with both *inform-basic* and *inform-proactive*. In the example proposed above to illustrate the dialogue act *inform-proactive*, sys provides the information directly requested by the user ("the company is located in Milan") as well as additional information (i.e. the company's email address).

3.1.2 Slots

A set of slots describes the relevant information we want to extract from dialogues in this specific domain. In our case each *slot* represents a specific attribute of the domain "job-offer". More specifically, we consider 14 domain-specific slots, described below:

- **age**:information referring to the age of the applicant or of the professional figure sought;
- **area**: sector of job position (e.g., *"I'd like to work in the advertising and communication area"*;
- **company-name**: name of the company or institution offering the job;
- **company-size**: company size based on the number of people who work there (e.g. *"I'd like to work in a big company"*);
- **contact**: contact information;
- **contract**: type of job contract offered or requested (e.g. *"part time"*);
- **degree**: degree or other qualification required or possessed by the applicant;
- **duties**: main tasks required by the job;
- **job-description**: title of the job position (e.g. *"web developer"*, *"receptionist"*);
- **languages**: knowledge of foreign languages required for the job or spoken by the user;
- **location**: location of the job or of the company;
- **past-experience**: user's previous work experiences;
- **skill**: skills requested for the job or possessed by the applicant;
- **other**: all the extra information related to the job-offer domain and not fitting other slots.

```
"turn_id": 17,
"usr": "Perfect. Could you tell me the name of the company or an address I can contact to get all the information I need?",
(...)
"turn_id": 20,
"sys": "I don't have the website's link. But I can give you the email address",

"async": [
    [
        "turn_ref",
        "turn_17"
    ]
],
```

Figure 2
An example of annotation of asynchronous messages.

All the semantically informative text fragments in dialogic turns are annotated with the dialogue acts and slots names. In addition to the domain-specific slots, the annotation schema also includes two general slots. The first one, **Global slot**, is used to mark the overall results of the dialogue and it can assume only two values, *positive* or *negative*, according to the outcome of the job interview. The label *positive* is used to express success in finding a useful job position, while the label *negative* is used in case of failure. Therefore, respect to the other slots, the Global slot refers not to the single utterances but to the entire dialogue. The second one, **Async**, is used to mark the

presence of asynchronous messages, which naturally occur in chat conversations. We consider asyncronous those overlapping utterances where the answer to a question is not immediate but comes in a later turn. When this phenomenon occurs, we mark as *async* the message where the speaker replies to the question, entering as value of the slot the number of the dialogic turn where the question was asked, as in the example in Figure 2.

3.2 Annotating JILDA

The annotation task we proposed is complex since all slot fillers are open classes and the values correspond to substrings extracted from text. The selection of these values was left to annotators' choices and therefore the boundaries of the selected text spans often differ, depending on the subjective choices made by the annotator.

JILDA and MTurk annotation process was supported by MATILDA, an open source tool specifically designed to annotate multi-turn dialogues, which was extended to support the management of collaborative annotation projects (Cucurnia et al. 2021). Each annotator is assigned subsets of the collection to annotate and can add/modify her own annotations without affecting the work of the others. The system takes care of persistence by storing in a database intermediate work of the annotators and offers management and monitoring capabilities to the project supervisor. The work of different annotators can be compared through a inter-annotator interface, which also supports the resolution of disagreements.

Annotating JILDA involved four annotators, who worked in pairs during two distinct annotation phases. Both JILDA and MTurk dialogues where annotated, thus building a dataset of over 750 fully annotated dialogues in the job search domain.

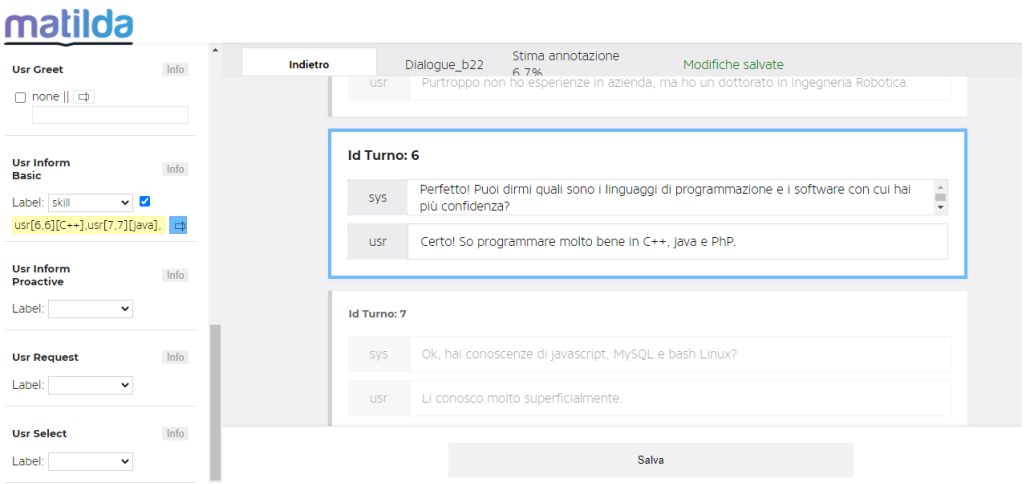

Figure 3
Dialogue annotation using MATILDA's interface

Figure 3 shows an example of dialogue annotation via the MATILDA's interface. Each dialogue, organised into dialogic turns, is shown in the middle of the interface screen. Each turn includes both system's and user's utterance. The panel on the left allows the annotator to select the relevant tags, filling the values of the slots through a text selection made directly from the input sentences. Besides the slot value, the position

in the sentence of the highlighted tokens is also stored. The annotated dialogues are then exported in json format, as shown in Figure 4.


```
"sys": "So, tell me, do you know any foreign languages?",
"sys_request": [
    [
        "languages",
        "sys[3,3][?],"
    ]
],
"turn_id": 5,
"usr": "Yes, I do know English (B2) and French (B2) and I have attended an A1 German course.",
"usr_inform_basic": [
    [
        "languages",
        "usr[4,4][English],usr[5,5][B2],usr[7,7][French],usr[8,8][B2],usr[14,14][A1],usr[15,15][German],"
    ]
]
```


Figure 4
Output of the annotated dialogue, in json format

4. Analysis

4.1 First Annotation Phase

The first annotation phase involved two annotators: one worked on the entire JILDA dataset, while the other annotated the Mturk collection. When this annotation was completed, we conducted a first analysis targeting the number of tokens and types per slot, in order to understand the frequency of use of the slots, their lexical variability and for each slot the size of the linguistic dictionary that can be extracted from JILDA and Mturk.

Table 2
Tokens and types extracted per slot during the first annotation phase

	tokens	types	Type/token ratio
age	92	27	0.29
area	873	447	0.51
company-name	464	107	0.23
company-size	392	238	0.60
contact	512	49	0.09
contract	987	170	0.17
degree	863	459	0.53
duties	1206	852	0.70
job-description	660	275	0.41
languages	795	142	0.17
location	1200	257	0.21
other	106	93	0.87
past-experience	588	463	0.78
skill	1287	659	0.51
Total	**10025**	**4238**	**0.42**

As shown in Table 2, the *type / token ratio* of the slots' values annotated in JILDA and Mturk is 0.42 on the average. These data suggest that the two datasets have a significant semantic variability and seem to effectively capture the linguistic variety of native speakers. On the other hand, a low type/token ratio can create difficulties in training an effective linguistic model, particularly when there is the need to generalise among slot classes. To overcome this problem, without losing the linguistic richness which is typical of JILDA, we introduced specific modifications and additional indications during the second annotation phase, as described in the next section.

In addition to analysing the vocabularies of both datasets and slots, we computed the number of proactive phenomena annotated. This is an interesting analysis to conduct, since it constitutes a measure of the complexity and naturalness of the data collected.

In JILDA 17.15% of dialogue acts were proactive, while in the MTurk dataset only 1.98%. This difference between JILDA and Mturk is undoubtedly due to the different data collection methodology used to build the two datasets: a template-based approach in the case of MTurk and a less rigid approach based on the Map Task methodology in the case of JILDA.

4.2 Second Annotation Phase

At the end of the first annotation phase, we noticed some critical issues. First of all, dialogue acts and slots were not linked. This means that an utterance could be marked with one (or more) acts but could lack of slots' values and, vice versa, selected slot values could pertain to different speech acts. Consequently, it was not possible to identify a posteriori which part of the text had been marked with a specific dialogue act. Moreover, as said before, the use of open classes for the slots has led to the production of a large vocabulary for both datasets, a possibly critical issue if the data are to be used to train a dialogue model.

In order to improve the quality of the annotation and to ensure greater consistency with the Multiwoz schema, we introduced the following adjustments in the configuration model and annotation guidelines:

- One or more slots were directly associated with one of the annotated dialogue acts, in accordance with Multiwoz's annotation schema.
- We asked annotators to include in the slot's selection the smallest informative part of an utterance. In this way, sentences like *"I would like to work as web developer"* were reduced to *"web developer"*.
- To avoid losing relevant information, in case of short confirmation or denial in a speaker's utterance, the referent of this speech act was made explicit, annotating as slot's value the relevant part of the text that appeared in the previous utterance. For example, if the system says *"I find a job offer as a nurse"* and the user says *"Ok, fine"*, the latter utterance is marked as usr-select (as dialogue act) + job-description (slot) + "nurse" (slot value).
- To comply with the Multiwoz schema, a request is always targeted to a specific slot, and the slot value is *"?"*.

Table 3
Types extracted per slot during the second annotation phase

	tokens	**types**	**Type/token ratio**
age	130	36	0.27
area	1472	331	0.22
company-name	556	96	0.17
company-size	732	149	0.20
contact	827	44	0.05
contract	1486	131	0.08
degree	1243	315	0.25
duties	1741	956	0.54
job-description	1362	425	0.31
languages	1085	60	0.05
location	1922	168	0.08
other	559	184	0.32
past-experience	882	244	0.27
skill	1994	570	0.28
Total	**15991**	**3709**	**0.23**

Following these changes to the guidelines, a second annotation phase was then realised. The work involved two different annotators, who equally shared the annotation work of JILDA and Mturk. This second annotation was more accurate and led to the creation of a more detailed dataset. Furthermore, from the analysis conducted after the annotation, it seems that the changes in the revised guidelines have actually led to a reduction of the corpus vocabulary, without however losing the lexical richness of the annotated data. Indeed, Table 3 shows that the vocabularies of the two datasets are still large, although the type/token ratio, which is 0.23, is lower than before (the type/token ratio of the previous annotation was 0.42).

Moreover, the number of proactive elements is still significant, with an overall percentage of 10.4% and this is a clear indicator of the naturalness and richness of the JILDA dataset with respect to MTurk. In fact, 12.7% of the dialogue acts in JILDA are proactive, while in MTurk we observe only 2.6% of proactive acts, also due to the different features of the dialogues.

4.3 Interannotator Agreement

In order to evaluate the quality of the annotated data, we calculated the inter-annotator agreement (IAA). We decided to compute the agreements between the two annotation rounds since annotators of both rounds worked on the same datasets and they had the same task, although the guidelines changed as described in 4.2. We computed the agreement in three different steps.

Firstly, we considered if there was an overlap between the text selected as slot value by the first annotator (A1) and the second one (A2). Indeed, it was important to consider if both the annotators recognised as ”informative” the same part of the utterance. We

decided to consider as an agreement also an approximated overlap. The example below shows two cases of accepted match, which is exact in the first example:

A1: ["usr-inform-proactive", "skill", **"bachelor's degree in engineering"**]
A2: ["usr-inform-basic", "degree", **"bachelor's degree in engineering"**]

and approximated in the second one.

A1: ["usr-inform-proactive", "skill", **"bachelor's degree in engineering"**]
A2: ["usr-inform-basic", "degree", **"degree in engineering"**]

From the 1725 strings identified by at least one of the annotators as informative, we identified **810** cases of agreement. By focusing on these overlapping values, we move on to consider whether the text fragments identified as informative were associated to the same slot by the annotators, as in the example:

A1: ["usr-inform-proactive", **"degree", "degree in engineering"**]
A2: ["usr-inform-basic", **"degree", "degree in engineering"**]

Finally, when there is a match both on values and on slots, we evaluated if there is an agreement also in the dialogue act, as in the example:

A1: [**"usr-inform-basic", "degree", "degree in engineering"**]
A2: [**"usr-inform-basic", "degree", "degree in engineering"**]

Using this approach, we computed three values for agreement: i.) the percentage of sub-string matches over the total number of selected values, ii.) the percentage of agreements in slot attribution over the total of matching sub-strings, and iii.) the percentage of agreements in dialogue-acts over the cases matching in both values and slots.

We computed the above agreement measures for JILDA and obtained the results shown in Table 4. We can observe that the agreement values are very low, as expected considering that changes made in the guidelines before the second round of annotation were substantial.

Table 4
IAA between first and second annotation on 10% of the dataset.

	Sub-strings	**Slot**	**Dialogue acts**
Cases	1725	810	714
Agreement	810	714	419
Accuracy	0.47%	0.88	0.58

To effectively evaluate the quality of the new annotation, we asked the two volunteers of the second phase to make a cross-annotation using a subset of JILDA, which corresponds to about 10% of the entire dataset. In this way we could evaluate if the workers had truly internalised the annotation scheme and had produced a consistent dataset. The new calculation of accuracy gives substantially higher values, as it can be seen from Table 5; this clearly proves that using the same guidelines annotators are able to create a consistent annotation of the dataset. In addition to the accuracy values, in this case we also computed Cohen's kappa both for dialogue acts and slots considering both the actual accuracy and the predicted accuracy. The results are extremely positive and are, respectively, 0.82 and 0.86. These values were computed on the basis of the the

confusion matrices between the two annotators reported in the Appendix. By looking at those matrices we can notice that, as slots are concerned, the two annotators often disagreed on the attribution to the slot *area* vs *degree* or *skill* vs *job-description*. In the attribution of slots to dialogue acts instead, most disagreements where associated, as expected, to the subtle distinction between *inform-basic* and *inform-proactive*.

Table 5
IAA between second and third annotation on 10% of the dataset.

	Sub-strings	Slots	Dialogue acts
Cases	1661	1230	1163
Agreement	1230	1163	911
Accuracy	0.73	0.87	0.84
Cohen's kappa	-	0.86	0.82

5. Grounding & Proactivity

The semantic annotations reported so far focused on slots related to the domain and to proactive dialogue acts. For what concerns the analysis of the proactivity in JILDA, we computed the number of labels used to mark information provided proactively by the speaker, as shown in Figure 5.


```
"sys": "Have you ever worked in public accounting?",
"sys_request": [
    [
        "past_experience",
        "sys[6,6][?],"
    ]
],
"turn_id": 8,
"usr": "No, but I can use Excel very well.",
"usr_deny": [
    [
        "past_experience",
        "usr[0,0][No],"
    ]
],
"usr_inform_proactive": [
    [
        "skill",
        "usr[5,5][Excel],"
    ]
]
```


Figure 5
Example of information provided proactively by the speaker.

As can be observed from Table 6, the number of proactive sentences, is quite high in JILDA, which constitutes a clear indicator of the naturalness of the data collected.

Although dialogues were not annotated with grounding phenomena, as exemplified in the introduction, we expect the JILDA dataset to include a substantial amount

Table 6
Number of proactive acts labelled in JILDA and MTurk.

	JILDA	**Mturk**
I annotation	2624	76
II annotation	1712	102
I ann % of proact. data	17.16%	1.98%
II ann % of proact. data	12.7%	2.6%

of instances of grounding for the fact that dialogues are natural and representative of unconstrained and cooperative human-to-human dialogues. In order to substantiate this claim with a quantitative analysis we can look at the presence of several patterns commonly associated with grounding expressions specific to this domain: expressions of confirmation, of misunderstanding and confusion, or requests for explanations.

Table 7
Grounding expressions in JILDA.

Pattern	**Instances**
capisco, capire, capito	284
ok	465
certo	402
certamente	188
chiaro, chiarire	15
d'accordo	115

Table 7 reports the number of instances associated to the corresponding patterns. This analysis is limited by the fact that manifestations of grounding expressed through questions are often hard to be distinguished from normal discovery questions about unknown features of the job offer or of the applicant profile.

Table 8
Grounding acts according to Traum (1999). DU stands for *Dialogue Units*.

Label	**Description**
initiate	Begin new DU, content separate from previous uncompleted DUs
continue	some agent adds related content to open DU
acknowledge	Demonstrate or claim understanding of previous material by other agent
repair	Correct (potential) misunderstanding of DU content
Request Repair	Signal lack of understanding
Request Ack	Signal for other to acknowledge
cancel	Stop work on DU, leaving it ungrounded and ungroundable

To give an idea of the progress of the grounding contribution within a dialogue, we have represented a portion of the JILDA dialogue presented in the Appendix as a state transition diagram, based on the model proposed in (Traum and Nakatani 2002). Using the grounding scheme proposed by Traum (see Table 8), the respective grounding acts have been identified for the first 16 turns of the dialogue, as shown in Table 9. It can be noted how *continue* and *acknowledge* constitute the core of the grounding behaviour. Particularly the applicant introduces new information (e.g., T9. *...should I specify my geographical preferences?*) only after the navigator has acknowledged, implicitly, the previous turn (T7. *letâĂŹs see immediately among the offers available what could fit best for you*).

Table 9
Grounding diagram for a portion of a JILDA dialogue.

Dialog. Turns	*initiate*	*continue*	*acknowledge*	*repair*	*Req. Repair*
T1	x				
T2			x		
T3			x		
T4		x			
T5		x			
T6		x			
T7			x		
T8			x		
T9		x			
T10			x		
T11			x		
T12			x		
T13		x			
T14			x		
T15					x
T16				x	

6. Conclusion and Future Work

We have presented JILDA, a corpus of annotated human-human goal-oriented dialogues related to the job-offer domain. Differently from other datasets, JILDA has been collected through map-task, a method allowing to acquire natural dialogues. As a result, JILDA dialogues exhibit both high linguistic variability and high presence of collaborative phenomena. Annotations take as a basis the MultiWOZ scheme but, differently from the latter, we annotate both user and system utterances, highlighting the dialogue acts describing the aim of the utterance, as well as slots specific to the JILDA job-offer domain. We presented a detailed analysis of the JILDA semantic annotations, showing that the new dataset contains a large amount of pragmatic phenomena, such as *proactivity* (i.e., providing information not explicitly requested) and *grounding*, which are both rarely investigated in current AI conversational agents based on neural architectures.

Given its innovative characteristics, JILDA has the potential to foster research in conversational AI toward really collaborative goal-oriented systems. To this end, we

intend to use JILDA to experiment neural dialogue state tracking and dialogue policy models able to reproduce both grounding and proactive interactions.

References

Balaraman, Vevake and Bernardo Magnini. 2020a. Investigating proactivity in task-oriented dialogues. In *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020*, volume 2769 of *CEUR Workshop Proceedings*, Bologna, Italy, March 1-3. CEUR-WS.org.

Balaraman, Vevake and Bernardo Magnini. 2020b. Proactive systems and influenceable users: Simulating proactivity in task-oriented dialogues. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2020)*, Virtually at Brandeis, Waltham MA, USA, July.

Balaraman, Vevake and Bernardo Magnini. 2021. Domain-aware dialogue state tracker for multi-domain dialogue systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:866–873.

Balaraman, Vevake, Seyedmostafa Sheikhalishahi, and Bernardo Magnini. 2021. Recent neural methods on dialogue state tracking for task-oriented dialogue systems: A survey. In Haizhou Li, Gina-Anne Levow, Zhou Yu, Chitralekha Gupta, Berrak Sisman, Siqi Cai, David Vandyke, Nina Dethlefs, Yan Wu, and Junyi Jessy Li, editors, *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2021)*, pages 239–251, Singapore and Online, July 29-31. Association for Computational Linguistics.

Budzianowski, Paweł, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium, October-November. Association for Computational Linguistics.

Chakravarty, Saurabh, Raja Venkata Satya Phanindra Chava, and Edward A. Fox. 2019. Dialog acts classification for question-answer corpora. In *Proceedings of the Third Workshop on Automated Semantic Analysis of Information in Legal Texts co-located with the 17th International Conference on Artificial Intelligence and Law (ASAIL@ICAIL)*, Montreal, QC, Canada, June.

Cheng, Jianpeng, Devang Agrawal, Hector Martinez Alonso, Shruti Bhargava, Joris Driesen, Federico Flego, Dain Kaplan, Dimitri Kartsaklis, Lin Li, Dhivya Piraviperumal, et al. 2020. Conversational semantic parsing for dialog state tracking. *arXiv preprint arXiv:2010.12770*.

Clark, H. Herbert and Susan E. Brennan. 1991. Grounding in communication. In L.B. Resnick, J.M. Levine, and S.D. Teasley, editors, *Perspectives on Socially Shared Cognition*. American Psychological Association, pages 127–149.

Clark, H. Herbert and F. Edward Schaefer. 1987. Collaborating on contributions to conversations. *Language Cognition and Neuroscience*, pages 19–41.

Cucurnia, Davide, Nikolai Rozanov, Irene Sucameli, Augusto Ciuffoletti, and Maria Simi. 2021. Multi-annotator multi-language interactive light-weight dialogue annotator. *In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 32 – 39, address=Online, April 19 – 23 2021.

Deriu, Jan, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *In Artificial Intelligence Review, 54, 755–810*.

Eric, Mihail, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 422–428, Marseille, France, May. European Language Resources Association.

Grice, Herbert P. 1975. Logic and conversation. In *Speech acts*. Brill, pages 41–58.

Henderson, Matthew, Blaise Thomson, and Jason D Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th annual meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272, Philadelphia, PA, USA, 18-20 June.

Henderson, Matthew, Blaise Thomson, and Steve Young. 2014. Word-based dialog state tracking with recurrent neural networks. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 292–299, Philadelphia, PA, USA, 18-20 June.

Hough, Julian, Iwan Kok, David Schlangen, and Stefan Kopp. 2015. Timing and grounding in motor skill coaching interaction: Consequences for the information state. In *Proceedings of the*

19th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2015), pages 86 – 94, Gothenburg, Sweden, August.

Hough, Julian and David Schlangen. 2017. It's not what you do, it's how you do it: Grounding uncertainty for a simple robot. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI 2017)*, pages 274–282, Vienna, Austria, March.

Kelley, J. F. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Trans. Inf. Syst.*, 2(1):26–41.

Levinson, Stephen C. 1983. *Pragmatics*. Cambridge University Press, Cambridge, U.K.

Louvan, Samuel and Bernardo Magnini. 2020. Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 480–496, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Mana, Nadia, Susanne Burger, Ronaldo Cattoni, Laurent Besacier, Victoria MacLaren, John McDonough, and Florian Metze. 2003. The NESPOLE! voIP multilingual corpora in tourism and medical domains. In *Proceedings of the 8th European Conference on Speech Communication and Technology (INTERSPEECH 2003)*, Geneva, Switzerland, September.

Mana, Nadia, Roldano Cattoni, Emanuele Pianta, Franca Rossi, Fabio Pianesi, and Susanne Burger. 2004. The Italian NESPOLE! corpus: a multilingual database with interlingua annotation in tourism and medical domains. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May. European Language Resources Association (ELRA).

Mrkšić, Nikola, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Vancouver, Canada, July. Association for Computational Linguistics.

Mushin, Ilana, Lesley Stirling, Janet Fletcher, and Roger Wales. 2003. Discourse structure, grounding, and prosody in task-oriented dialogue. *DISCOURSE PROCESSES*, 35:1–31, 01.

Ramadan, Osman, Paweł Budzianowski, and Milica Gasic. 2018. Large-scale multi-domain belief tracking with knowledge sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 432–437, Melbourne, Australia July 2018.

Shah, Pararth, Dilek Hakkani-Tür, Bing Liu, and Gokhan Tür. 2018. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 41–51, New Orleans - Louisiana, June. Association for Computational Linguistics.

Shan, Yong, Zekang Li, Jinchao Zhang, Fandong Meng, Yang Feng, Cheng Niu, and Jie Zhou. 2020. A contextual hierarchical attention network with adaptive objective for dialogue state tracking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6322–6333, Online, July. Association for Computational Linguistics.

Sucameli, Irene, Alessandro Lenci, Bernardo Magnini, Maria Simi, and Manuela Speranza. 2020. Becoming JILDA. In Johanna Monti, Felice Dell'Orletta, and Fabio Tamburini, editors, *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020*, pages 409–414, Bologna, Italy (Online), March 1-3, 2021. CEUR-WS.org.

Traum, David R. 1999. Computational models of grounding in collaborative systems.

Traum, David R. and Christine H. Nakatani. 2002. A two-level approach to coding dialogue for discourse structure: Activities of the 1998 DRI working group on higher-level structures. In *Towards Standards and Tools for Discourse Tagging*.

Wen, Tsung-Hsien, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M. Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain, April. Association for Computational Linguistics.

Wu, Chien-Sheng, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819, Florence, Italy, July. Association for Computational Linguistics.

Zang, Xiaoxue, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for*

Conversational AI, ACL 2020, pages 109–117, Online, July.
Zhou, Li and Kevin Small. 2019. Multi-domain dialogue state tracking as dynamic knowledge graph enhanced question answering. *ArXiv*, abs/1911.06192.

Appendix

Sample of JILDA dialogues

An illustrative example of goal-oriented dialogue from JILDA.

1. Navigator: *Ciao! Sono Chiara e sarÃš il tuo navigator. Come posso aiutarti?.*

2. Applicant: *Ciao Chiara. Mi chiamo Marta e sono alla ricerca di un lavoro*

3. Navigator: *Ciao Marta, spero di poterti aiutare. Mentre cerco fra le offerte disponibili, puoi dirmi qualcosa sul tuo lavoro ideale?*

4. Applicant: *Mi sono laureata da pochi mesi in Legge*

5. Applicant: *Mi piacerebbe poter imparare dai professionisti del mio ambito, quindi poter essere assunta, anche per un tirocinio, in uno studio di avvocatura o notarile sarebbe per me il massimo*

6. Applicant: *Mi rendo conto che come ambiente sia sovraccaricato di offerta e che i posti aperti sono molto pochi perÃšâĂę*

7. Navigator: *Mai perdere la speranza Marta e io sono qui proprio per aiutarti a cercare l'offerta migliore per te! Vediamo subito fra le offerte disponibile cosa potrebbe fare al caso tuo..*

8. Applicant: *Grazie, sei molto gentile*

9. Applicant: *Nel frattempo potrei specificarti le mie preferenze a livello geografico? Potrebbero aiutarti nel targetizzarmi meglio*

10. Navigator: *SÃň, perfetto! Grazie*

11. Applicant: *Attualmente vivo in Toscana: sono disponibile a trasferirmi in altre regioni ma anche allâĂŹestero non ho problemi di mobilitÃă o limiti da questo punto di vista*

12. Navigator: *Potrei avere due offerte che mi piacerebbe proporti. Entrambe riguardano tirocini post-laurea, uno come assistente capocommessa in una azienda edile a Pistoia, e l'altra come allievo direttore a Milano presso Compass. Ho pensato a te perchÃÍ entrambi gli impieghi riguardano incarichi gestionali e la tua laurea potrebbe essere molto utile in questi contesti.*

13. Navigator: *Uno dei due ti sembra piÃź interessante e vuoi che te lo descriva per primo?*

14. Applicant: *Devo essere sincera: il primo non penso che possa fare al caso mio. Potresti descrivermi il secondo lavoro?*

15. Applicant: *Non riesco a capire bene che cosa significhi "allievo direttore"*

1. Navigator: *Hello! I'm Chiara and I'll be your navigator. How can I help you?.*

2. Applicant: *Hi Chiara. My name's Marta and I'm looking for a job*

3. Navigator: *Hi Marta, I hope I can help you. While I search for available offers, can you tell me something about your dream job?*

4. Applicant: *I graduated in Law few months ago.*

5. Applicant: *I'd like to learn from experts in my area and be hired, even for an internship, in a law firm or notary's would be great for me.*

6. Applicant: *I realize that this sector is overloaded with requests and that there are very few places open, butâĂę*

7. Navigator: *Never give up hope, Marta, I'm here to help you find the best offer available for you. Let's see immediately among the offers available what could fit best for you ..*

8. Applicant: *Thanks, you're very kind*

9. Applicant: *In the meantime, should I specify my geographic preferences? They could help you target me better*

10. Navigator: *Yes, perfect! Thank you*

11. Applicant: *At the moment I live in Tuscany: I'm available to move to other regions and even abroad I don't have mobility problems or limitations from this point of view*

12. Navigator: *I may have two offers that I would like to propose to you. Both involve postgraduate internships, one as an assistant prime contractor in a construction company in Pistoia, and the other as a junior director in Milan at Compass. I think they could be ok for you because both jobs involve management assignments and your degree could be very useful in these contexts.*

13. Navigator: *Does one of them seem more appealing? And do you want me to describe it first?*

14. Applicant: *I have to be honest: I don't think the first one is right for me. Could you describe the second job for me?*

15. Applicant: *I can't quite understand what "junior director" means*

16. Navigator: *Certo! Le principali mansioni legate a questo impiego riguardano la pianificazione del budget e del conto economico dell'azienda. Il settore ÃÍ quello alimentare quindi si tratta di compilare ordini e derrate alimentari, oltre che garantire la sicurezza sul lavoro e quella alimentare.*

17. Navigator: *Compiti gestionali sarebbero sicuramente al centro del lavoro.*

18. Navigator: *Ti sembra piÃź chiaro? Posso dirti altro?*

19. Applicant: *Capisco. Mi sembra interessante*

20. Applicant: *Sai se lâĂŹazienda offre formazione? A chi per esempio come me non ha un grande background economico ma ha fatto esami allâĂŹUniversitÃă di Economia*

21. Navigator: *Trattandosi di un tirocinio postlaurea direi che la formazione sarÃă una componente importante.*

22. Applicant: *Capisco. CâĂŹÃÍ una deadline per fare domanda?*

23. Applicant: *Sto considerando anche altre posizioni aperte in parallelo*

24. Navigator: *No, tranquilla, puoi inviare domanda quando preferisci. Per ora la posizione ÃÍ aperta.*

25. Applicant: *Sai quanti candidati prendono? Per prepararmi al meglio alla selezione*

26. Navigator: *Per ora ho solo questa offerta pubblicata da questa azienda. Penso che si tratti di un solo posto aperto.*

27. Navigator: *E non si indica il numero di domande accettate.*

28. Applicant: *Capisco. Potresti darmi il contatto dellâĂŹazienda? In modo tale da approfondire e mettermi in contatto diretto con loro*

29. Navigator: *Certo! Eccolo info@azienda.com*

30. Applicant: *Ti ringrazio per lâĂŹorientamento. Ti lascio il mio contatto, nel caso in cui saltino fuori posizioni aperte nel mio campo: martamarta@gmail.com*

31. Navigator: *perfetto! Se ci fossero novitÃă ti contatterÃš!*

32. Navigator: *Spero di esserti stata comunque utile.*

33. Applicant: *Molto, A presto e buona giornata*

34. Navigator: *Buona gioranta anche a te!*

16. Navigator: *Sure! The main tasks related to this job concern the planning of the budget and the income statement of the company. The area is the food sector so it's a question of filling orders and foodstuffs, as well as guaranteeing work and food safety.*

17. Navigator: *Management tasks would certainly be the core of the work.*

18. Navigator: *Is it more clear now? Can I tell you more?*

19. Applicant: *I see. It seems interesting*

20. Applicant: *Do you know if the company offers training? For example, for those who, like me, don't have a background in economics but took some exams at the University in Economics*

21. Navigator: *Since this is a post-graduate internship I would say that training will be an important component.*

22. Applicant: *I see. Is there an application deadline?*

23. Applicant: *I'm considering other open positions in parallel*

24. Navigator: *No, don't worry, you can apply whenever you like. The position is open for now.*

25. Applicant: *Do you know how many candidates they accept? To better prepare myself for the selection*

26. Navigator: *For now I only have this offer published by this company. I think it's just one open position.*

27. Navigator: *And the number of applications accepted is not indicated.*

28. Applicant: *I see. Could you give me the company's contact? This way I can take a closer look and contact them directly*

29. Navigator: *Sure! It's info@azienda.com*

30. Applicant: *Thank you for the assistance. I'll give you my contact info, in case open positions arise in my field: martamarta@gmail.com*

31. Navigator: *perfect! If there is any news I will contact you!*

32. Navigator: *I hope I have been helpful, anyway.*

33. Applicant: *Very, see you soon and have a good day*

34. Navigator: *Good day to you too!*

Confusion matrix on slots and dialogue acts

Confusion matrix between annotator A and annotator B on 10% of the JILDA dataset in classifying overlapping text spans into slots.

A \ B	age	area	comp name	comp size	contact	contract	degree	duties	job descr	lang	location	none	other	exp	skill	Sum
age	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**12**
area	1	150	0	0	0	0	35	1	10	0	1	0	5	3	9	**215**
comp_name	0	0	78	0	4	0	0	0	1	0	2	0	1	0	0	**86**
comp_size	0	1	0	83	0	1	0	1	0	0	1	0	4	0	0	**91**
contact	0	0	3	0	104	0	0	0	0	0	0	0	1	0	0	**108**
contract	0	1	0	2	0	169	0	0	6	0	0	0	6	2	2	**188**
degree	0	19	0	0	0	0	202	0	1	1	0	0	3	1	1	**228**
duties	0	4	0	0	0	0	0	217	11	2	1	0	0	0	8	**243**
job_descr	0	14	1	0	1	8	0	12	124	0	1	0	2	1	1	**165**
languages	0	0	0	0	0	0	6	0	0	134	0	0	11	0	1	**152**
location	0	0	2	2	1	1	0	2	3	0	225	0	1	0	0	**237**
none	0	1	0	0	0	0	0	0	0	0	1	294	1	3	0	**300**
other	0	3	0	1	0	0	3	3	1	10	0	0	2	0	1	**24**
experience	0	6	0	0	0	1	1	2	1	0	0	0	3	68	6	**88**
skills	0	4	0	0	0	0	3	15	1	2	0	0	2	11	285	**323**
Sum	**13**	**203**	**84**	**88**	**110**	**180**	**250**	**253**	**159**	**149**	**232**	**294**	**42**	**89**	**314**	**2460**

Figure 6
Agreement on slots

Confusion matrix between annotator A and annotator B on 10% of the JILDA dataset in classifying slots into dialogue acts.

A \ B	sys deny	sys greet	sys inform basic	sys inform proactive	sys request	sys select	usr deny	usr greet	usr inform basic	usr inform proactive	usr request	usr select	**Sum**
sys_deny	11	0	1	1	0	0	0	0	0	0	0	0	**13**
sys_greet	0	134	0	0	0	0	0	0	0	0	0	0	**134**
sys_inform_basic	1	0	505	57	1	1	0	0	0	0	0	0	**565**
sys_inform_proactive	0	0	56	34	1	0	0	0	0	0	0	0	**91**
sys_request	0	0	2	0	217	0	0	0	0	0	0	0	**219**
sys_select	0	0	1	0	0	97	0	0	0	0	0	0	**98**
usr_deny	0	0	0	0	0	0	21	0	0	5	0	0	**26**
usr_greet	0	0	0	0	0	0	0	160	0	0	0	0	**160**
usr_inform	0	0	0	0	0	0	1	0	376	105	2	0	**484**
usr_inform_proactive	0	0	0	0	0	0	0	0	87	129	0	0	**216**
usr_request	0	0	0	0	0	0	0	0	0	2	126	0	**128**
usr_select	0	0	0	0	0	0	0	0	1	0	0	12	**13**
Sum	**12**	**134**	**565**	**92**	**219**	**98**	**22**	**160**	**464**	**241**	**128**	**12**	**2147**

Figure 7
Agreement on dialogue acts