Becoming JILDA

Irene Sucameli
Dipartimento di Informatica
Università di Pisa
irene.sucameli@phd.unipi.it

Alessandro Lenci
Dipartimento di Filologia, Letteratura, Linguistica
Università di Pisa
alessandro.lenci@unipi.it

Bernardo Magnini
Fondazione Bruno Kessler
Trento
magnini@fbk.eu

Maria Simi
Dipartimento di Informatica
Università di Pisa
simi@di.unipi.it

Manuela Speranza
Fondazione Bruno Kessler
Trento
manspera@fbk.eu

Abstract

English. The difficulty in finding useful dialogic data to train a conversational agent is an open issue even nowadays, when chatbots and spoken dialogue systems are widely used. For this reason we decided to build JILDA, a novel data collection of chat-based dialogues, produced by Italian native speakers and related to the job-offer domain. JILDA is the first dialogue collection related to this domain for the Italian language. Because of its collection modalities, we believe that JILDA can be a useful resource not only for the Italian research community, but also for the international one.

Italiano. *Negli ultimi anni l'utilizzo di chatbot e sistemi dialogici è diventato sempre più comune; tuttavia, il reperimento di dati di apprendimento adeguati per addestrare agenti conversazionali costituisce ancora una questione irrisolta. Per questo motivo abbiamo deciso di produrre JILDA, un nuovo dataset di dialoghi relativi al dominio della ricerca del lavoro e realizzati via chat da parlanti nativi italiani. JILDA costituisce la prima collezione di dialoghi relativi a questo dominio, in lingua italiana. Per gli aspetti metodologici e la modalità di raccolta dei dati, riteniamo che una simile risorsa possa essere utile ed interessante non solo per la comunità di ricerca italiana ma anche per quella internazionale.*

1 Introduction

Chatbots and spoken dialogue systems are now widespread; however, there is still a main issue connected to their development: the availability of training data. Finding useful data to train a system to interact as human-like as possible is not a trivial task. This problem is even more critical for the Italian language, where only few datasets are available. To supplement this deficiency of data, we decided to develop **JILDA** (*Job Interview Labelled Dialogues Assembly*), a new collections of chat-based mixed-initiative, human-human dialogues related to the job offer domain. Our work offers different elements of novelty. First of all, it constitutes, to the best of our knowledge, the first dialogue collection for this domain for the Italian language. Moreover, our dataset was not built using a Wizard of Oz approach, usually adopted in the realization of dialogues. Instead, we used an approach similar to the Map Task one, as we will describe in the next section. This allowed us to obtain more complex, mixed-initiative dialogues.

2 Background

Few dialogic datasets are available for Italian, including the NESPOLE dialogues related to the tourism domain (Mana, 2004), QA datasets related to the movie or the customer care domains (Bentivogli, 2014), and a recent dataset derived from the translation of the English SNIPS (Castellucci, 2019). However, the resources currently available are still limited and, to the best of our knowledge, none of the existing ones is related to the domain of job-offer. For what concerns the English language, although there are more dialogic resources that can be used to train conversational agents (Lowe, 2015; Yu, 2015; El Asri, 2017; Budzianowski, 2018; Li, 2018), as far as we know there are no relevant and freely accessible datasets related to job-matching. Moreover, these

Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-2769/paper_69.pdf

datasets usually record simplified conversations, which do not represent the effective complexity that characterises human-human interactions. To fill this gap, we decided to produce a new dialogic dataset for the job domain, for the Italian language. To collect data representative of the linguistic naturalness of native speakers, we had to detect the best approach to fulfil our aim.

The WoZ approach. One of the common approaches used to build full-scale datasets is Wizard of Oz (WoZ) (Kelley, 1984), where a human (the wizard) covers the role of the computer within a simulated human-computer conversation. The other participants in the conversation, however, are not aware that they are talking to a human rather than a conversational system (Rieser, 2008). This method has pros and cons: it may allow to collect conversations written in natural language in a short time (Wen, 2017); however, the dialogues built in this way may not record the noisy conditions experienced in real conversations (e.g. repetitions, errors) and do not show much variation from the syntactic and semantic point of view (Budzianowski, 2018). Due to the limitations of WoZ, we decided to adopt other methods to build our dataset. The first method used in an initial phase of experimentation, was the template-based approach.

The template-based approach. In this solution, it is asked to a volunteer to paraphrase template dialogues using natural language in order to create a simulated dialogue (Shah, 2018). We experienced this modality during an initial experimental phase, in which we used templates for creating task-oriented dialogues. In this first experiment, as previously done by Shah et al. (Shah, 2018), we used Amazon Mechanical Turk[1] and we asked Italian native speakers to cover the role of both the computer and the user, paraphrasing templates of dialogues between a recruiter and a job seeker. We proposed three different templates, with 15-20 recruiter-user interactions each and, to ensure greater lexical variety, we inserted some random variables into the templates (for example, user's skills and the type of job requested). With this experimental set up, we built a first dataset of 220 dialogues. However, despite the attempts to ensure linguistic variety, we noticed that in the MTurk dataset the conversation was strongly guided by the templates provided and that the dialogues were little diversified from a lexical point of view.

The Map Task approach. To overcome the limits of the WoZ and of the template-based approach, and to produce a set of mixed-initiative dialogues which reflect the naturalness typical of human-human interaction, we decided to organise a new experiment. In this second phase of experimentation, we used as guideline the methodology adopted for the Map Task experiment (Brown, 1984), in which two participants collaborate to achieve a common purpose. For example, Anderson et al. adopted the Map Task to build the HCRC Corpus (Anderson, 1991), a corpus of dialogue recordings and transcriptions. Realized in a similar way, but for the italian language, there is the CLIPS[2] corpus, a dataset containing speech recordings.

In Anderson's Map Task, one speaker (the Instruction Giver) has a route marked on the map while the other speaker (the Instruction Follower) has the map without the route and, talking with the Instruction Giver, has to reproduce the route. However, the two maps are not identical and the participants have to discover how they differ.

In our experiment, the two parts involved had to collaborate in a conversation to find the best match between job-offer and candidate profile. The participants covered the role of the *navigator*[3], who had a set of possible job offers, and of the *applicant*, who was provided with a job profile to impersonate (a short CV). While in the HCRC Map Task the two parts had to interact in order to figure out the route on the blind map, in this case the two participants had to chat to find the best job-offer match possible for both parts. In the next section, both the framework and the set up of our experiment are described in detail.

3 Experimental setup

To create the JILDA dialogues collection for job-offer, we asked 50 Italian native speakers to simulate a conversation between a "navigator" and an applicant. At the end of the experiment, all the volunteers received an economical reward for their participation. We randomly assigned to 25 volun-

[1] Available here: `https://www.mturk.com/`

[2] Available here: `http://www.clips.unina.it/it/corpus.jsp`

[3] The navigator plays a role similar to the recruiter's one, who is in charge of reviewing candidate's skills and past experiences in order to find a suitable job.

teers the role of navigator, providing 5 job offers each. The other 25 volunteers had to pretend to be applicants and describe themselves on the basis of the information contained in a curriculum we provided. The navigators' goal was to help applicants to find a job offer (among the offers available) best suited to their curriculum and interests by asking questions. Applicants, on the other side, had to interact with the navigator describing the skills and competencies included in their curricula.

Similarly to the Map Task framework, the two parties had to collaborate in order to reach their goal and were engaged in creating a mixed initiative spontaneous dialogue without a strict guidance. Navigators and applicants were free to lead the conversation as they preferred; in fact, we did not use any dialogue template (although we provided some examples) and both applicants and navigators were allowed to ask questions to their interlocutor, in order to reach the best possible match between applicant's needs and the job offers available to the navigator. The only compulsory requirements we imposed to participants was to converse only about topics related to the experiment. In addition to this, we provided as guideline an indicative length of 15/20 (overall) utterances per dialogue.

Both navigators and applicants were not allowed to interact with the same interlocutor twice. Each navigator interacted with 21 different applicants and, in a similar way, each applicant had to interact with 21 navigators. With this strategy we wanted not only to obtain dialogues as linguistically diversified as possible, but also to ensure that navigators with different offers interacted with applicants with different curricula and needs.

To make the navigator interact with the applicant, we used the Slack platform[4], which allowed the volunteers to interact with each other in an easy way, maintaining anonymity through the use of nicknames. Moreover, it allowed us to monitor multiple conversations at the same time and to easily download the dialogues' output in a json format suitable for the future annotations. Neither the applicants nor the navigators knew with whom they had to chat.

We asked the volunteers to realise 21 chat-based dialogues distributed in five days, so they had to produce 4 or 5 dialogues per day.

[4]Available at `https://slack.com/intl/en-it/`

4 Results and Discussion

At the end of the experiment, we collected 525 chat-based, mixed initiative dialogues [5]. In order to have a first evaluation of the data produced, we asked our volunteers to assess the quality of the dialogues. More specifically, we asked to evaluate the degree of naturalness, the linguistic variety of the dialogues (Table 1), and the difficulties detected in the experiment (Table 2). Among the 50 participants, 29 completed the evaluation questionnaire. The results obtained are reported below.

Rating Scale	Realism	Linguistic variety
1 (very low)	0%	0%
2	7%	14%
3	14%	55%
4	62%	21%
5 (very high)	17%	10%

Table 1: *Evaluation of the degree of realism and linguistic variety of JILDA dialogues.*

Rating scale	Difficulty in understanding	Difficulty in the description
Very low	0%	0%
Medium	17%	48%
Very high	83%	52%

Table 2: *Evaluation on the degree of difficulty in understanding the interlocutor's requests and in describing the job offers/CV available.*

The volunteers' evaluation is in line with what can be observed directly from the dialogues. In fact, from a preliminary analysis, the dialogues produced exhibit a good linguistic variety and capture complex phenomena of the Italian language, such as co-reference. Since they are task oriented dialogues, the data follow a certain pattern of questions/answers but, within this common structure, the navigator-applicant interaction varies in an extremely interesting way. For instance, we noticed the presence of asynchronous messages with respect to the context, as shown in the example reported in Appendix A. This is due to the fact that users have the tendency to type fast while they are chatting, and this may lead to overlapping messages, were the answer to a question is not immediate but comes in a later turn. Furthermore,

[5]Both JILDA and MTurk datasets are available here: `http://dialogo.di.unipi.it/jilda/`

applicants do not passively answer to navigators but they often take the initiative, formulating questions and proactively giving unsolicited information. Comparing JILDA's dialogues with MTurk's ones, it is clear that JILDA's dialogues are more complex and semantically diversified.

	MTurk	**JILDA**
# dialogues	220	525
avg turns/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
# proactive/intent	1.97%	17.30%
# proactive/sentences	1.46%	12.70 %

Table 3: *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 of which includes the same number of tokens as MTurk*

A first analysis, for which we also used Profiling-UD (Brunato, 2020) and UDPipe (Straka, 2017), highlights differences of the new dataset with respect to the previous one [6] such as:

- **lexical variability.** As shown in Tab.3, JILDA has a greater lexical variability, which is extremely useful if the dataset is used to train new models. In fact, considering the whole dataset, JILDA has more tokens and types. Even more importantly, by selecting subsets of JILDA with the same number of tokens as MTurk, it is possible to verify that, on the average, JILDA's lexical richness is higher (see the lemma and type/token ratio).

- **syntactic complexity.** With respect to the MTurk dataset, JILDA includes more subordinates and longer chains of dependencies, which is an indication of more complex sentences. In fact, the analysis conducted with Profiling-UD (Brunato, 2020) shows for JILDA a higher percentage of subordinate propositions (51.46% against 39.87% in MTurk) and longer chains of embedded subordinate clauses (18.35% of the chains are long 2 or more in JILDA, 12.48% in MTurk).

- **dialogue naturalness.** The naturalness of JILDA's dialogues partially emerged in the first evaluation conducted with the participants in the experiment (Table 1-2). In addition to this, Table 3 shows that JILDA contains a high number of proactiveness phenomena, which are significant in highlighting the complexity of a dialogue and its collaborative nature. In particular, JILDA contains a higher number of proactive intents, both in terms of percentage over the total number of intents and over the number of sentences. [7] This shows that our volunteers did not merely answer their interlocutor by providing the strictly required information, but rather on their own initiative provided additional information, which made the dialogues more natural and complex.

The annotation of the dialogues is now in progress in order to offer to the scientific community not only a new set of dialogues for the Italian language but also, and above all, a richly annotated dataset. The annotation will take as a basis the notation of Multiwoz, which is becoming a standard in dialogue datasets (Budzianowski, 2018). However, although in Multiwoz only user's turns are annotated, we decided to annotate both applicant's and navigator's utterances, since we noticed that both utterances convey important and useful information. The preliminary analysis of the data presented here will be deepened once the annotation is complete. To support the annotation work of the JILDA dataset, we modified an open source dialogue annotation tool, LIDA, in collaboration with its developers (Collins, 2019). Specifically, we extended this tool to 1) allow support for multiple annotators working at the same project, 2) manage multiple annotation styles and metadata information, 3) manage different collections of dialogues and 4) simplify the annotation interface, improving the user experience. Both the new release of the LIDA Multi-user annotation tool and the JILDA annotated dataset will be made available to the scientific community.

[6] It is worth to highlight that the differences between the two resources are primarily related to the methods used for data collection and not to the platforms used.

[7] Proactive intents were explicitly annotated for this count.

5 Conclusion

In this paper we presented JILDA, a novel dataset of chat-based, mixed-initiative dialogues built for the Italian language and related to the job-offer domain. This new resource has been built adopting an experimental approach based on the Map Task experiment. This has allowed us to collect mixed-initiative data which represent effectively the naturalness which is typical in the human-human interaction. The JILDA dataset, which includes 525 dialogues, is in the process of being completely annotated with dialogue acts and entities related to this specific domain. For the annotation of those dialogues we are using our own extension of LIDA. The annotated dialogues will then be used to train a conversational agent. Thanks to this new resource, our goal is to allow an agent chat with the user in a natural and human-like way.

Acknowledgments

This work has been endorsed by AILC (Italian Association for Computational Linguistics).
We thank Carla Congiu, Clara Casandra and Davide Cucurnia, students of Digital Humanities at the University of Pisa, for annotation work on JILDA and for contributing to the development of the annotation tool.

References

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S. and Weinert, R., 1991. The HCRC Map Task Corpus. *Language and Speech*, 34, pp. 351-366.

Bentivogli, L., Magnini, B., 2014. An Italian Dataset of Textual Entailment Graphs for Text Exploration of Customer Interactions. In *Proceedings of the first Italian Computational Linguistics Conference*.

Brown, G., Anderson, A., Yule, G., Shillcock, R., 1984. *Teaching talk: Strategies for production and assessment*. Cambridge University Press.

Brunato D., Cimino A., Dell'Orletta F., Montemagni S., Venturi G., 2020. Profiling-UD: a Tool for Linguistic Profiling of Texts". In *Proceedings of 12th Edition of International Conference on Language Resources and Evaluation (LREC 2020)*, pp. 11-16 May, 2020, Marseille, France.

Budzianowski, P. Tsung-Hsien, W., Bo-Hsiang, T. Casanueva, I., Ultes, S., Ramadan, O., Gašić, M., 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 (EMNLP)*, pp. 5016-5026.

Castellucci, G., Bellomaria, V. and Favalli, A., Romagnoli, R.,, 2019, Multi-lingual Intent Detection and Slot Filling in a Joint BERT-based Model. In ArXiv abs/1907.02884.

Collins, E., Rozanov, N., Zhang, B. 2019 LIDA: Lightweight Interactive Dialogue Annotator. In *Proceedings of the 2019 EMNLP and the 9th IJCNLP* (System Demonstrations), pp. 121–126.

Dell'Orletta, F., Montemagni, S., Venturi, G., 2011 READ-IT: assessing readability of Italian texts with a view to text simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies (SLPAT 2011)*, Association for Computational Linguistics, pp. 73-83.

El Asri, L., Schulz, H., Sharma, S., Zumer, J., Harris, J., Fine, E., Mehrotra, R., Suleman, K. 2017. Frames: A Corpus for Adding Memory to Goal-Oriented Dialogue Systems. In arXiv:1704.00057

Kelley, J.F. 1984. An iterative design methodology for user-friendly natural language office information applications. In *ACM Transactions on Information Systems (TOIS)*, 2(1), pp. 26-41.

Li, R., Kahou, S.E., Schulz, H., Michalski, V.,Charlin, L., Pal, C. 2018. Towards Deep Conversational Recommendations, In Advances in Neural Information Processing Systems 31 (NIPS 2018), pp. 9748-9758.

Lowe, R., Pow, N., Serban, I. and Pineau J. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *Proceedings of the SIGDIAL 2015 Conference*, pp. 285-294.

Mana, N., Cattoni, R., Pianta, E., Rossi, F., Pianesi, F., and Burger, S. 2004. The Italian NESPOLE! Corpus: a Multilingual Database with Interlingua Annotation in Tourism and Medical Domains. In *Proceedings of 4th International Conference LREC*.

Rieser, V., Lemon, O., 2008. Learning Effective Multimodal Dialogue Strategies from Wizard-of-Oz Data: Bootstrapping and Evaluation. In *Proceeding of ACL-08:HLT*, pp. 638-646.

Shah, P., Hakkani–Tür, D., Liu, B., Tür, G., 2018. Bootstrapping a Neural Conversational Agent with Dialogue Self-Play, Crowdsourcing and On-Line Reinforcement Learning. In *Proceeding NAACL-HLT 2018*, pp. 41-45.

Straka, M. and Straková, J., 2017 Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 88-99 Vancouver, Canada, August 2017.

Wen, T.-H., Vandyke, D., Mrksic, N., Gasic, M., Rojas-Barahona, L.M., Su, P.-H., Ultes, S., Young, S., 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*, vol. 1, pp. 438–449.

Yu, Z., Papangelis, A., Rudnicky, A.I., 2015. TickTock: A Non-Goal-Oriented Multimodal Dialog System with Engagement Awareness. AAAI Spring Symposia.

Appendix A

Example of asynchronous message in JILDA

Navigator: *Cercano persone che si occupino sia di gestire la comunicazione pubblicitaria del cliente attraverso il web, che di interagire direttamente con la clientela.*

Applicant: *Quanto tempo dura il periodo di formazione?*

Navigator: *Questo significa che abilità di comunicazione sono essenziali in questo lavoro*

Applicant:

Navigator: *L'annuncio non fornisce informazioni circa la durata del contratto, mi dispiace*

Appendix B

Example of dialogue from Mturk

sys: *Salve e benvenuto alla Recruiter Top, io sono Tony.*

usr: *Buongiorno Tony, mi chiamo Giorgio e sono alla ricerca di un lavoro come traduttore.*

sys: *Bene, mi dica qualcosa in più su di lei; attualmente lavora o studia? e quali sono le sue competenze?*

usr: *Mi sono appena laureato in lingue e letterature straniere, nello specifico con conoscenza di inglese, spagnolo e francese*

sys: *E per quanto riguarda esperienze lavorative?*

usr: *Ho lavorato 2 anni in una casa editrice*

sys: *in che senso ha lavorato per 2 anni in una casa editrice?*

usr: *Sono stato viceredattore per una casa editrice locale.*

sys: *Ok, e per quello che riguarda le lingue straniere? essendo laureato in lingue...*

usr: *bhe si..ovviamente ho una buona conoscenza di francese , inglese e spagnolo*

sys: *Che contratto e tipologia di azienda sta cercando o valuterebbe?*

usr: *Propenderei per un tempo determinato in una azienda all'estero*

sys: *al momento si cerca per la sede di Gais figura professionale che si occupi della corrispondenza telefonica e scritta con i nostri clienti spagnoli e portoghesi e di organizzare eventi di marketing, potrebbe essere interessato?*

usr: *Si, mi dia i dettagli e lo valuterò. Grazie e arrivederci*

sys: *Arrivederci e buona fortuna.*

Example of dialogue from JILDA

sys: *Ciao, sono il tuo Navigator di oggi, mi chiamo Mattia. Posso aiutarti in qualche modo?*

usr: *Buongiorno Mattia, mi chiamo Valentina e sto cercando un lavoro a tempo determinato.*

sys: *Ciao Valentina, puoi dirmi qualcosa in più sugli studi che hai fatto?*

usr: *Certamente! Mi sono laureata tre anni fa in Lingue e Letterature straniere.*

sys: *Ottimo, hai già avuto esperienza lavorativa in passato o sarebbe il tuo primo lavoro?*

usr: *Ho già avuto un'esperienza lavorativa, perché per due anni ho lavorato come guida museale.*

sys: *Ti è mai capitato di lavorare a progetti con bambini, durante questi due anni?*

usr: *Quando lavoravo per il museo non ho mai affrontato dei progetti specifici riguardanti i bambini. Ho però fatto da guida a delle scolaresche.*

sys: *Ho qui un annuncio riguardo la possibilità di fare assistenza scolastica a minori con disabilità, dalle scuole d'infanzia alle superiori. Pensi che ti piacerebbe provare qualcosa del genere?*

usr: *Sarebbe un'esperienza interessante, ma non credo di avere le competenze necessarie. Preferirei rimanere nel campo dei musei o, in generale, in quello dei luoghi turistici.*

sys: *Al momento non ho annunci per posti disponibili in campo turistico o museale, mi dispiace. Data la tua laurea in Lingue però, vorrei proporti un annuncio di CHANEL Cordination S.r.l., sono alla ricerca di una stagista da affinacare alla Responsabile Qualità Prodotto referente per l'Italia.*

usr:

sys: *dovresti occuparti principalmente di Monitoraggio del database dei prodotti delle collezioni. Gestione dei contatti con i fornitori locali ed esteri.Archiviazione e consultazione dei Test di laboratorio e supporto della responsabile nella preparazione di presentazioni in PPT e nelle traduzioni della reportistica nelle lingue in inglese e francese*

usr: *Mi interesserebbe molto. Dove si trova l'azienda?*

sys: *La sede dell'azienda è a Milano, quindi probabilmente dovrai spostarti lì se non abiti già in zona,*

usr: *Non sarebbe un problema spostarmi. Il lavoro é a tempo pieno o a tempo parziale?*

sys: *Non è specificato nell'annuncio, so solo che si tratta di un tirocinio/stage. Probabilmente è un cosa da discutere in fase di colloquio direttamente con loro*

usr: *Ok grazie.*

sys: *Puoi contattare direttamente l'azienda a questo indirizzo e-mail info@azienda.com*

usr: *Perfetto, grazie mille! :)*

sys: *Figurati, buona fortuna per il lavoro!*

usr: *Grazie, buona giornata! :)*