Proceedings of the EURALI Workshop @LREC2020, pages 51–60
Marseille, 20 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

Dialogue Act and Slot Recognition in Italian Complex Dialogues

Irene Sucameli*, Michele De Quattro*, Arash Eshghi, Alessandro Suglia**, Maria Simi***
*Department of Computer Science, University of Pisa
*School of Mathematical and Computer Sciences, Heriot-Watt University
irene.sucameli@phd.unipi.it, m.dequattro@studenti.unipi.it,
a.eshghi@hw.ac.uk, as247@hw.ac.uk, simi@di.unipi.it

Abstract

Since the advent of Transformer-based, pretrained language models (LM) such as BERT, Natural Language Understanding (NLU) components in the form of Dialogue Act Recognition (DAR) and Slot Recognition (SR) for dialogue systems have become both more accurate and easier to create for specific application domains. Unsurprisingly however, much of this progress has been limited to the English language, due to the existence of very large datasets in both dialogue and written form, while only few corpora are available for lower resourced languages like Italian. In this paper, we present JILDA 2.0, an enhanced version of a Italian task-oriented dialogue dataset, using it to realise a Italian NLU baseline by evaluating three of the most recent pretrained LMs: Italian BERT, Multilingual BERT, and AlBERTo for the DAR and SR tasks. Thus, this paper not only presents an updated version of a dataset characterised by complex dialogues, but it also highlights the challenges that still remain in creating effective NLU components for lower resourced languages, constituting a first step in improving NLU for Italian dialogue.

Keywords: Dialogue systems, Italian dataset, NLU

1. Introduction

The field of Natural Language Processing (NLP) was transformed when Vaswani et al. (2017) presented their self-attention-based, Transformer model for representation or embedding of Natural Language strings, with Devlin et al. (2019) then releasing BERT, a large scale pretrained LM, showing that new state of the art results could be obtained in many canonical NLP tasks just by fine-tuning with one additional task-specific output layer. This *transfer learning* methodology forms the basis of the most important component in dialogue systems today: Natural Language Understanding (NLU). Moreover, it has also been applied to our problem of interest in this paper: that of Dialogue Act Recognition (DAR, e.g. Chakravarty et al. (2019)) combined with Slot Recognition (SR), tasks which aim to evaluate how well a system classifies the dialogue acts (i.e the goal of the speaker's utterance) and the slots (i.e the informative elements which have to be extracted in order to understand and fulfil the speaker's goal) of a sentence.

Much of the progress above has, however, been limited to the English language due largely to the unavailability of high quantities of language corpora in other languages. For example, in comparison to English, Italian stands as a lower resourced language and, with few exceptions (Mana et al., 2004; Castellucci et al., 2019; Sucameli et al., 2020), there is currently a paucity of dialogue datasets available with appropriate Dialogue Act & Slot annotations for training effective NLU models. Large scale multilingual models do exist (e.g. Multilingual BERT), but it is as yet unclear how these models *transfer* to the NLU tasks of DAR & SR. One important reason for this uncertainty is that nearly all existing, large-scale LMs have been trained on open domain, written language, whereas dialogue is known to be very different from text or written language: dialogue is highly context-dependent, is replete with fragments (Fernández and Ginzburg, 2002; Purver et al., 2009), ellipsis (Colman et al., 2008) & disfluencies (Shriberg, 1996; Hough, 2015), and is highly domain-specific (Eshghi et al., 2017). Noble and Maraev (2021) provide evidence for this, showing that pretrained BERT does not transfer well for the DAR task without being fine-tuned on the target dialogues. In this paper, we focus on NLU for dialogue systems in Italian. We present and use an enhanced version of the JILDA corpus (Sucameli et al., 2020) – one of the very few Italian dialogue datasets in the public domain – to evaluate three of the most recent pretrained LMs on the DAR & SR tasks: Multilingual BERT (Devlin et al., 2019), Italian BERT (Schweter, 2020), and AlBERTo (Polignano et al., 2019).

2. Related work

2.1. BERT for dialogue NLU

Ever since the advent of the Transformer model, BERT (Devlin et al., 2019) has become the de facto standard for the DAR and SR tasks, and has seen success in many dialogue domains in the English language (Mehri et al., 2019; Ribeiro et al., 2019; Chakravarty et al., 2019; Bao et al., 2020). For these tasks, a *transfer learning* method is employed using BERT, which uses a multi-layer bidirectional transformer to embed the input text. In such approaches, BERT is used as the pretrained encoder, whose one or more hidden layers are fed to additional output layer(s) or classifiers and fine-tuned on specific in-domain NLU datasets. Considering the effectiveness of such a transfer learning approach for dialogue, Noble and Maraev (2021) show, interestingly, that the pretrained model isn't of much use without fine-tuning on target dialogue data.

In this paper, we study the usefulness of three different versions of BERT as the pretrained language model, and evaluate their performance in the DAR & SR tasks on the JILDA 2.0 dataset, a new updated version of the collection of mixed-initiative, human-human dialogues in Italian, and in the 'job offer' domain originally presented in Sucameli et al. (2020).

2.2. Dialogue Datasets

Annotated dialogue corpora are at the core of the capacity to learn dialogue models. Among human-human corpora, it is certainly worth citing the **ReDial dataset** (Li et al., 2018), which includes 10,000 human-human recommendation dialogues collected via Amazon Mechanical Turk.

The **Twitter Corpus** (Ritter et al., 2010) also belongs to this category, with 1.3 million post-reply pairs extracted from Twitter; **The Ubuntu Dialogue Corpus** (Lowe et al., 2015) is another with a large amount of unstructured dialogues used to train dialogue systems without any NLU annotations (see e.g. Lowe et al. (2017)). There is also a number of human-machine dialogue corpora: this includes the **DSTC1** dataset (Williams et al., 2013), a popular task-oriented dataset released in conjunction with the Dialog State Tracking Challenge; and the **Frames** dataset (Asri et al., 2017), which studies user's decision-making behaviour. Finally, belongs to this category MultiWOZ, a collection of dialogues built with a Wizard of Oz (WoZ) approach.

MultiWoZ (Budzianowski et al., 2018) is one of the most influential dialogue datasets with a recent 2.3 version released (Han et al., 2021) which addresses some annotation errors of the original. MultiWoZ 2.0 contains 10,438 dialogues collected using the WoZ approach and which cover various domains, such as restaurant and hotel search, taxi and hospitals. Thanks to the frequent updates of the dataset, MultiWoZ constitutes an important benchmark for Natural Language Understanding.

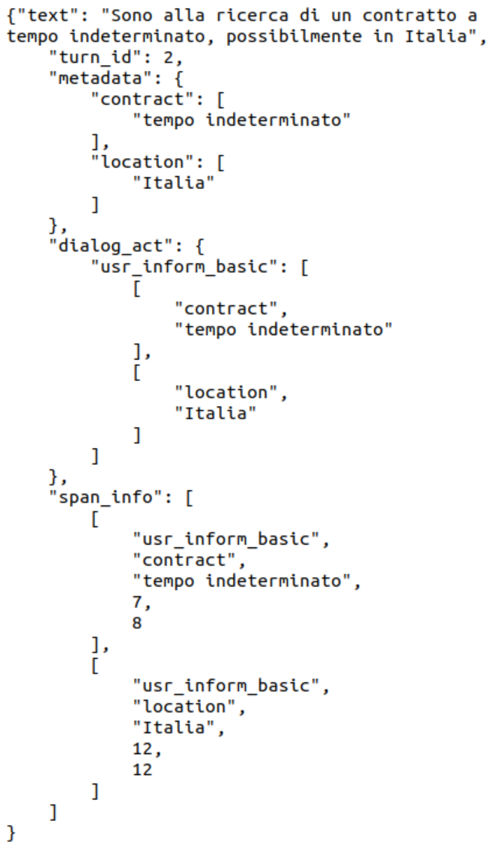
2.3. Italian Dialogue Datasets

In comparison to English, in which there are numerous dialogue datasets available (see Li et al. (2018; Lowe et al. (2017; Budzianowski et al. (2018; Liu et al. (2021) among many others), Italian is a lower resourced language: more specifically, there is currently a paucity of dialogue datasets available with appropriate Dialogue Act and Slot/Named Entity annotations for training effective NLU models. Among the few collections available are the **NESPOLE** dataset (Mana et al., 2004) in the tourism domain; the **SNIPS** dataset (Castellucci et al., 2019) – derived through translation from English; and the newly released **JILDA** dataset (Sucameli et al., 2020) which we use for our experiments in this paper.

3. The JILDA dataset

JILDA (Sucameli et al., 2020) is a collection of complex human-human dialogues realised in Italian, and in the 'job offer' domain. This dataset includes 745 mixed-initiative dialogues collected in an experiment which involved 50 Italian native speakers and was inspired by the Map-task methodology (Brown et al., 1984), in which two participants collaborate to achieve a common purpose (in this case, the realization of a task-oriented dialogue).

The produced resource consists of 17,889 utterances and a total of 263,104 tokens, characterised by great linguistic variability and syntactic complexity; indeed, the dataset presents, on average, 17 turns per dialogue with more of the 51% percentage of subordinate propositions (an example of JILDA dialogues is reported in Appendix). Furthermore, the datasets includes dialogues with linguistic phenomena that are often not contained or considered in the collections of dialogues, such as proactive and grounding phenomena. These phenomena, typical of human-human conversations, confirm, together with the evaluations made, the naturaleness of the dialogues produced.



```
{"text": "Sono alla ricerca di un contratto a
tempo indeterminato, possibilmente in Italia",
    "turn_id": 2,
    "metadata": {
        "contract": [
            "tempo indeterminato"
        ],
        "location": [
            "Italia"
        ]
    },
    "dialog_act": {
        "usr_inform_basic": [
            [
                "contract",
                "tempo indeterminato"
            ],
            [
                "location",
                "Italia"
            ]
        ]
    },
    "span_info": [
        [
            "usr_inform_basic",
            "contract",
            "tempo indeterminato",
            7,
            8
        ],
        [
            "usr_inform_basic",
            "location",
            "Italia",
            12,
            12
        ]
    ]
}
```

Figure 1: Example of a JILDA annotated dialogue.

JILDA has been annotated with the DAs and slots reported in Table 1, using MATILDA, an open source tool created to annotate multi-turn dialogues (Cucurnia

et al., 2021), following the annotation scheme of MultiWOZ 2.0 (Budzianowski et al., 2018). Budzianowski et al. (2018) used a set of 13 Dialogue Acts (such as: inform, greet, request, reqmore, not found) and 23 slots to annotate dialogues referred to 7 domains (restaurant, hotel, attraction, taxi, train, hospital, police)[1].
JILDA annotation schema includes 12 dialogue acts (DA) and 14 slot types; Figure 1 shows an example of JILDA annotation schema, while Table 1 shows the distribution of different dialogue acts and slots in the JILDA dataset. In addition to this, the dataset enjoys high inter-annotator agreement ($\kappa = 0.86$ for DAs; $\kappa = 0.82$ for Slots)(Sucameli et al., 2021).
Together, all these data highlight the complexity of JILDA, and indicate that creating effective NLU models for this data is likely to be challenging. In what follows, we evaluate three different NLU models (DAR+SR) on these dialogue datasets.

DA	Occur.	Slots	Occur.
usr-greet	3222	age	130
usr-deny	1257	area	1472
usr-select	890	company-name	556
usr-inform-basic	8665	company-size	732
usr-inform-proac.	3335	contact	827
usr-request	2940	contract	1486
sys-greet	2918	degree	1243
sys-deny	759	duties	1741
sys-select	1868	job-description	1362
sys-inform-basic	6736	languages	1085
sys-inform-proac.	1590	location	1922
sys-request	6494	other	559
		past-experience	882
		skill	1994

Table 1: JILDA DA and Slot occurences.

4. JILDA 2.0

In order to be able to make a comparison between our Italian NLU model and the model based on MultiWOZ 2.1 (Han et al., 2021), one of the main benchmarks for English NLU, we decided to upgrade the current version of JILDA, realising **JILDA 2.0**.
JILDA 2.0, now available on Github [2], constitutes an updated version of the resource, implemented with design choices compliant with MultiWOZ 2.1. Specifically, we made some improvements to the annotations of the first version of the dataset, as illustrated below:

1. correction of inferred annotations. For example, the following sentence[3]:

   ```
   sys: "Si tratta appunto di un
   tirocinio post-laurea, nel
   settore pubblicitario presso
   una azienda pisana"
   ```

   was annotated as:

   ```
       "sys_inform_basic": [
           ["location","Pisa"]]
   ```

   In this case, ”Pisa”, although cannot be found in the sentence, was inferred from the adjective ”pisana”.

2. resolution of turns' annotations which were marked using dialogue acts and slots related to the next turn, due to an incorrect use by annotators of the MATILDA tool. For example[4]:

   ```
   sys: "Cercano persone che
   si occupino di gestire la
   comunicazione pubblicitaria
   del cliente attraverso il web"
       "sys_inform_basic": [
       ["duties", "gestire la
       comunicazione pubblicitaria"],
       [ "skill","abilità di
       comunicazione"]]
   sys: "Questo significa che
   abilità di comunicazione sono
   essenziali"
   ```

   To resolve this error, the annotation has been referred to the correct turn and text spans have been updated.

3. adjustment of tokens boundaries. For example, the sentence[5]:

   ```
       sys: "Il candidato (...)
       ha inoltre il compito di
       gestire le comunicazioni
       per il cliente e le
       informazioni su richiesta
       dell'ospite"
   ```

   was annotated with:

   ```
       "sys_inform_basic": [
           ["duties","informazioni
           su richiesta dell'ospit"]]
   ```

[1] In MultiWOZ 2.0 not all the DAs and slots are used over all the domains.

[2] http://github.com/IreneSucameli/JILDA

[3] Translation: ” *It is a post-graduate internship, in the advertising sector at a Pisan company.*”

[4] Translation: ”*They are looking for people who manage the customer's advertising communication via web.” and ”This means that communication skills are essential.*”

[5] Translation: ”*The candidate (...) has also to manage the communications and information for the customer, if requested by the guest.*”

Here, the token ”ospit” cannot be found in the utterance since the final vowel is missing. This kind of error depends on the tool used for the annotation, which initially allowed to select the words' range based on the single characters, instead of the entire token. In JILDA 2.0 these errors have been fixed by inserting the correct token in the span.

4. resolution of annotations which embed information from previous turns. In MultiWOZ 2.1 slots and dialogue acts are annotated and extracted turn by turn, disregarding information coming from previous or following turns. For this reason we decided to conform to the MultiWOZ standard by removing the occurrences of this annotation from JILDA, which instead presents a great number of annotations which relies on information referring to previous turns. In fact, together with the changes described in the second point, a total of 882 occurrences have been removed. With this change, the resource has been more aligned with our reference model. An example can help to understand better the changes made. In the conversation below[6] the word ”da remoto” does not appear in the user's sentence, since the speaker is reffering to an implicit subject (here, the ”remote working”). Therefore, the information annotated in the user turn derives from the system's turn:

```
sys: "Ti piacerebbe lavorare da
remoto?",
usr: "Si, potrebbe andare bene!"
    "usr_inform_basic": [
    [ "location","da remoto"]]
```

To conform to MultiWOZ 2.1, we decided to remove annotations referring to previous turns from the turn's span; still, the extracted information remained stored in the metadata. We decided to maintain the information since we think it could be interesting, in future works, to use specialised tag-sets in order to effectively capture relevant linguistic inferences, similarly to what was done by Bentivogli et al. (2010).

The updated dataset, produced as a result of the changes illustrated above, was then used to evaluate three of the most recent pretrained LMs on the Dialogue Act and Slot recognition tasks. The experimental results are reported in the next sections.

5. Experimental Setup

5.1. Models

Our experiments were conducted within ConvLab-2[7], an open-source multi-domain end-to-end dialogue system platform realised by Zhu et al. (2020). We chose this tool in order to have results comparable to the ones produced by Han et al. (2021) with the ConvLab-2's BERTNLU module. This module, which was used for our experiments as well, is based on a pretrained BERT to which it adds on top two Multi-Layer Perceptrons (MLPs), one for dialogue act classification and another for slot tagging, as shown in Figure 2. Here, the Transformer model is called at different times within the same cycle. The number of layers depends on the pretrained LM used. For each sentence, it is called twice with the indicated inputs and outputs, and also produces a pooled representation of the context. Then, the Slot Classifier produces as many outputs as the words in the sentence, while the DAR returns a score on the different DA values.

In BERTNLU all those dialogue acts which appear in the utterances are converted using BIO tags, a common tagging format for tagging tokens in chunks (Ramshaw and Marcus, 1995). We used BERTNLU combined with three different language models available on Hugging Face: **bert-base-italian-xxl-cased** [8](Schweter, 2020), **bert-multilingual-cased** [9](Devlin et al., 2019) and **AlBERTo**[10](Polignano et al., 2019).

	bert-italian-xxl	bert-multil.	AlBERTo
Voc. Size	32K	119K	128K
Source	OPUS, OSCAR and Wikipedia	Wikipedia	TWITA

Table 2: Comparison of vocabulary size of the LMs.

The first one is trained on Wikipedia, on the OPUS corpus[11] (which includes - among the other data - transcripts of spoken language and subtitles) and on the Italian part of the OSCAR corpus[12], which consists of raw web pages. The second one is trained with the top 100 languages from Wikipedia, including Italian. Since the size of Wikipedia varies from language to language, and to avoid under-representation of lower resourced languages, in the multilingual version of BERT, high-resource languages (like English) are under-sampled, while lower resourced languages are over-sampled.

Finally **AlBERTo** (Polignano et al., 2019) is a BERT LM for the Italian language, trained on 200M tweets with a vocabulary size of 128k. AlBERTo replicates the BERT stack and it is trained using masked language modelling loss only since the authors remove the next sentence prediction loss because tweets don't have a notion of sequence of sentences like in documents.

[6] Translation: sys:”*Remote working would be ok for you?*” usr:”*Yes, it would be fine.*”

[7] https://github.com/thu-coai/ConvLab-2

[8] https://github.com/dbmdz/berts

[9] https://github.com/google-research/bert

[10] https://github.com/marcopoli/AlBERTo-it

[11] https://opus.nlpl.eu/

[12] https://oscar-corpus.com/

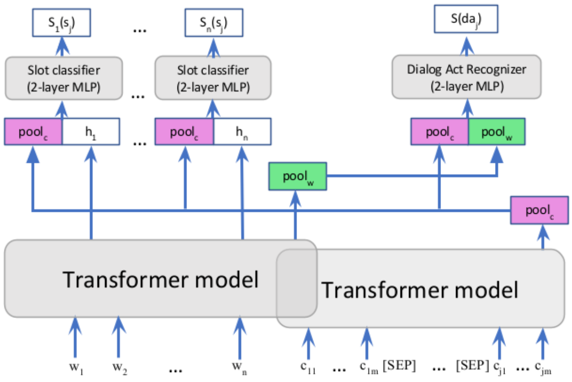

$S_1(s_j)$
...
$S_n(s_j)$
$S(da_j)$
Slot classifier (2-layer MLP)
...
Slot classifier (2-layer MLP)
Dialog Act Recognizer (2-layer MLP)
$pool_c$
h_1
...
$pool_c$
h_n
$pool_c$
$pool_w$
$pool_w$
$pool_c$
Transformer model
Transformer model
w_1 w_2 ... w_n
c_{11} ... c_{1m} [SEP] ... [SEP] c_{j1} ... c_{jm}

Figure 2: BERTNLU architecture. The Transformer models produce two types of pools, one for the words (w) and another for the contexts (c). These pools are sent to the Slot Classifier and the Dialogue Act Recognizer. There are as many Slot Classifiers as there are words, while for the Dialogue Act is produced a single distribution of probability on the different values.

5.2. Hyper-parameters

We used the JILDA 2.0 dataset to finetune & evaluate the above-mentioned models on the DAR & SR tasks, taking the 80% of the data for training (596 dialogues) & 20% for testing and validation (respectively, 75 and 74 dialogues). The hyper-parameter tuning procedure is described in Appendix. After fixing the hyper-parameters, we trained each model and computed average scores for Precision, Recall and F1 Score. Since JILDA used MATILDA's tokenizer, while ConvLab-2 adopts spacy and bert tokenizer, we decided to standardize the different tokenizations. Thus, it was decided to apply the spacy tokenizer [13] to JILDA 2.0 annotated data.

We also decided to unify the classification of the DA *inform-basic* and *inform-proactive*, since these two acts express the same intent, which could be expressed proactively (e.g. the speaker autonomously provided unsolicited information) or not; with a view to distinguishing the types of dialogues acts, it was therefore reasonable to consider them as a unique act. Moreover, even without this distinction, it is still feasible the comparison with the English benchmark since in MultiWOZ 2.1 there is not the difference between proactive and required dialogue acts.

In order to quantify how well each pretrained encoder – bert-base-italian, bert-multilingual & AlBERTo – encodes the target JILDA 2.0 dialogues, i.e. how well it transfers, we evaluated each model in two training conditions: 1) **end-to-end**, where the weights of the underlying encoder model were finetuned together with the task-specific DAR & SR layers; and 2) **frozen-lm** where the weights of the encoder layers were frozen with only the task-specific layers fine-tuned.

6. Results & Discussion

6.1. The end-to-end condition

Table 3 shows the averaged results obtained for the three models in the end-to-end condition. The overall results record the cases in which both the DAs and the slots in a sentence have been correctly predicted.

		bert-ita	bert-multi	AlBERTo
Acts	Prec.	81.55	**82.85**	79.74
	Rec.	**75.36**	70.41	70.66
	F1	**78.33**	76.12	74.92
Slots	Prec.	**71.65**	68.06	70.78
	Rec.	**71.27**	66.99	65.60
	F1	**71.46**	67.52	68.09
Overall	Prec.	**74.20**	71.66	73.13
	Rec.	**72.38**	67.92	66.97
	F1	**73.28**	69.74	69.92

Table 3: Values of Precision, Recall and F1 Scores in the `end-to-end` condition.

Analysing the performance of the models reported in Table 3, it can be firstly observed that the monolingual models perform better than the multilingual one. This proves that using LM in line with the language of the training data helps to reach better results in the recognition and classification of dialogue acts and slots. Nevertheless, the F1 score difference between the multilingual and monolingual BERT models is low enough to affirm that the first model is not less effective than the monolingual ones. This shows that at least the Italian language is represented well within the multilingual BERT model.

Among the three models, the best performing one definitely appears to be **bert-ita-xxl**. Comparing the monolingual models (bert-ita-xxl vs. AlBERTo) we noticed that bert-ita shows a superior performance than AlBERTo, which, however, has a larger vocabulary than the first one; in fact, the first one is originally trained on 81GB of data and 32K terms, while the second one consists of 191GB of raw data and a vocabulary of 128K terms. This demonstrates that LMs pre-trained on data similar to dialogues are able to gain better results than those trained on textual documents, regardless of the size of their dataset. Indeed, despite its size, AlBERTo is pretrained on Italian tweets, which tend to have a simplified structure compared to that of the JILDA dialogues used in our training. On the other hand, bert-ita-xxl is based on pre-training data that includes syntactically longer and semantically richer sentences (such as data from Wikipedia and OSCAR corpus), as well as transcripts of spoken conversation and subtitles (from the OPUS corpus), which present a syntactic and semantic structure close to that of the JILDA dialogues.

The results achieved are good if we consider that they were obtained using complex training dialogues. In fact, if we compare the results obtained by bert-ita (our best model) combined with JILDA 2.0 with those ob-

[13] `https://spacy.io/models/it`

tained by bert-base trained with MultiWOZ 2.1(Eric et al., 2020), it is possible to notice that the performance achieved by the Italian model is interesting. The comparison between the two datasets is feasible, although they differ in the dialog domain and in the size of the collected data, since they use the same architecture for training the NLU model for DAR and SR tasks. In fact, MultiWOZ 2.1 (Eric et al., 2020), which deals with some annotation errors of the previous version of the dataset, introduces an additional annotation for both user's and system's side and the resulting dataset is used to train, via ConvLab-2, the BERTNLU module for the DAR and SR tasks (Han et al., 2021). The results reported in Han et al. (2021) were obtained under similar conditions to ours (e.g. the `context` and the `fine-tune` hyper-parameters were set as `true`), and were evaluated using the same metrics. For all these reasons it was possible to compare the results obtained training the models with JILDA 2.0 with those reported in Han et al. (2021).

Datasets	F1 (Slot/DA/Both)
JILDA 2.0	71.46/78.33/73.28
MultiWOZ 2.1	81.18/88.34/83.77

Table 4: Performance of BERTNLU with JILDA and MultiWOZ 2.1.

	JILDA 2.0	MultiWOZ 2.1
Dialogues	745	10.438
Tokens	263.104	1.490.615
Ontologies' entries	5.779	2.111

Table 5: Comparison of JILDA 2.0 and MultiWOZ 2.1 in terms of dataset size and lexical vocabulary.

Table 4 shows how the F1 scores achieved by JILDA, although inferior to those of MultiWOZ, are not only reasonable but also very positive, if we consider that our model was trained using a dataset which is much smaller and, at the same time, extremely rich from a lexical point of view, as shown in Table 5. In fact, JILDA has far fewer dialogues and tokens but the number of values extracted from the ontology (which includes the lexical vocabulary of each slot) is over twice, sign of the linguistic richness of the Italian dataset.

Furthermore, compared to the original JILDA data, the improvements made to the new JILDA 2.0 version and described in Sections 4 & 5.1, allowed to increase the overall F1 score of the models trained under the end-to-end condition by almost 50 scores. This shows that the changes realised actually helped the NLU models to perform better.

Therefore, from the analysis of the results obtained, it is possible to state that the NLU model trained on our dataset shows convincing performances such as to be proposed a new benchmark for the Italian NLU.

6.2. The frozen-lm condition

Table 6 shows the averaged Precision, Recall & F1 Score values obtained in the `frozen-lm` condition where the weights of the encoder stack were frozen during training and only the task-specific heads fine-tuned.

		bert-ita	bert-multi	AlBERTo
Acts	Prec.	82.26	**96.00**	80.13
	Rec.	32.01	10.57	**54.51**
	F1	46.09	19.05	**64.66**
Slots	Prec.	70.15	63.80	**72.23**
	Rec.	**55.34**	48.26	50.22
	F1	**61.87**	54.96	59.25
Overall	Prec.	72.02	65.44	**74.34**
	Rec.	49.05	38.10	**51.38**
	F1	58.35	48.16	**60.77**

Table 6: Values of Precision, Recall and F1 Score recorded for the three models without fine-tuning the language model encoder stack.

Comparing Table 3, which shows the performance of the fine-tuned models, with Table 6, it is clear that fine-tuning the weights of the encoder model together with the task-specific DAR SR layers allows to gain better values. The results above are in line with those found by (Noble and Maraev, 2021) and highlight the importance of fine-tuning pre-trained encoders. Interestingly however, comparing the performance of the three models, when the fine-tune parameter is set to false, the one which performs better is AlBERTo. We believe that this is due to the data and vocabulary size used in the original training; in the absence of fine-tuning it seems that the model with more pre-training data obtains better performances.

7. Error Analysis

Having computed the F1 scores of the three models, we conducted an error analysis in order to verify which acts and slots were recognised more easily and which with more difficulties. To this end, we calculated the accuracy for the recognition of dialogue acts and slots and for each of the models. This measure is often used to evaluate NLU models and for intent detection task (Mohamad Suhaili et al., 2021), which is similar to our DAR and SR tasks.

	bert-ita	bert-multi	AlBERTo
DA Acc.	**78.25**	76.03	74.84
Slot Acc.	**71.46**	67.57	68.08

Table 7: Averaged accuracy in DAR and SR tasks.

Table 7 reports the overall accuracy computed per each model, while Figures 3 and 4 represent the accuracy of each DA and slot and for each model.

Analysing the accuracy of each DA (Figure 3), we noticed that *inform* had the highest values, while *greet* the lowest, probably due to the number of representation in the dataset of these acts. In fact, as illustrated in Table 1, some DAs and slots are higher represented than other; the higher is their representation in the dataset, the more accurate the models' classification is.

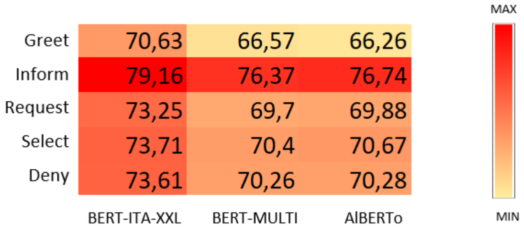


Figure 3: DAR accuracy in monolingual/multilingual BERT and AlBERTo.

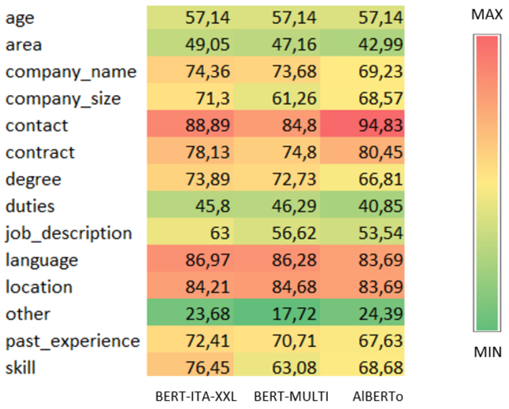


Figure 4: SR accuracy in monolingual/multilingual BERT and AlBERTo.

Regarding the classification of slots, it seems that the models had more difficulty with those slots which share lexical entries, as shown in Figure 4. When calculating slot accuracy, predicted slots were considered correct when they were used to classify the same part of text (ie the slot's value) marked by the true (ie gold) slots.

Excluding *other*, which constitutes a particular slot as it allows to mark all information which cannot be marked in another class, the slots whose recognition accuracy is less than 50.00 were only *area* and *duties*. The difficulty in recognizing these slots may be due to the fact that they presents larger and more open lexical vocabularies than those of slots such as *contract*, *contact* or *language*. In fact, for example, the lexical vocabulary of *duties* includes 985 entries, while *language* has 240 entries and *contact* less than 70.

Another aspect to take into consideration is the potential sharing of semantic contexts and syntactic structures. This means that a word, depending on the context in which it is found, could be annotated using multiple slots. Indeed, vocabulary overlapping between slots is a common phenomenon in JILDA.

For example, in the first sentence of Fig. 5[14] the text span can be annotated both with the slot *area* and with *degree*, due to the vocabulary overlap between these two slots. Similarly, in the second sentence, "insegnamento" can be considered both as a work's type or area, depending on the connotation we want to give to the term.

Cerco lavoro nel mio campo di studi. Mi sono laureato alla triennale in economia e marketing a Torino.

True label	area	economia e marketing
Predicted label	degree	economia e marketing

Al momento non abbiamo offerte di insegnamento.

True label	job_description	insegnamento
Predicted label	area	insegnamento

Figure 5: Overlap of slots' lexical vocabularies.

L'azienda ha la sua sede nella provincia di Pisa.

True label	location	provincia di Pisa
Predicted label	location	Pisa

Sono una neo-laureata in scienze sociali.

True label	degree	neo-laureata in scienze sociali
Predicted label	degree degree	neo-laureata scienze sociali

Figure 6: Text fragment selection errors. In the first example a part of the text is missing, while in the second one the relevant information is split in two slots.

Analysing the errors produced by the three models, it was also noted how, in some cases, even when the models correctly identified the relevant part of the text for the slots, they cut the informative text fragment, thus producing False Positive or False Negative. In the sentences in Figure 6 [15], for example, the model correctly recognised "Pisa", "neo-laureata" and "scienze sociali" as informative, annotating them with the correct act and slot. However, since the gold label included a larger text fragment, these predictions were considered as false by the model itself.

[14] Translation sentence n.1: " *I am looking for a job in my field of study. I graduated in Economics and marketing in Turin.*"
Sentence n.2: "*At the moment we don't have any teaching offers.*"

[15] Translation sentence n. 1: " *The company has its headquarters in the province of Pisa.*"
Sentence n.2: "*I recently graduated in social sciences*".

The analysis and the discussion conducted point out that creating effective NLU components for dialogue systems in domains grounded in data as linguistically rich & complex as JILDA remains a challenge. Therefore, starting from the values presented in Tab. 3, we propose in the future to further investigate the DAR and SR tasks for NLU Italian models, training the models with different recurring neural networks in order to achieve even a better performance.

8. Conclusion

In this paper we presented JILDA 2.0, an updated version of the Italian dataset collecting dialogues in the job application domain. In order to realise a NLU baseline trained with JILDA 2.0 that was comparable with the MultiWOZ 2.1 benchmark, we evaluated three recent pretrained LMs, namely Italian BERT, Multilingual BERT and AlBERTo. We fine-tuned and tested these models on the Dialogue Act Recognition and Slot Recognition tasks which are good proxy tasks for how well and under what training conditions these models are able to effectively encode dialogue semantics.

Our results showed that: (1) comparing the monolingual and the multilingual models, the first type resulted to be more able to obtain a better performance when specialised on an Italian dialogic dataset; (2) the size of the dataset used in the original training of the LM has less impact on the results than the type of data used in the original training; in fact, it was recorded a better performance for bert-ita-xxl, whose vocabulary is smaller than the one of AlBERTo but includes data which have linguistic features closer to those of the JILDA dialogues; (3) the multilingual BERT model performs only slightly worse than the monolingual model, highlighting the relative effectiveness of the multilingual model for the Italian language; and (4) fine-tuning the pretrained encoder is important, especially when the target data are dialogues that differ in many important ways from written data.

Furthermore, in comparison with the model trained on MultiWOZ 2.1, our NLU model presents convincing performances such as to constitute a new benchmark for the Italian NLU.

Our work demonstrates not only the issues related to the training of NLU models on lower resourced language, but, more importantly, constitutes a starting point for working on Italian models, specifically pretrained on dialogic dataset like JILDA. For future work, we will try to further refine the JILDA dataset and expand its annotation, in order to align the resource with the current version of MultiWOZ 2.3. Finally, we would like to introduce a Bidirectional LSTM in the BERTNLU architecture in order to improve the results of the current NLU module.

9. Bibliographical References

Asri, L. E., Schulz, H., Sharma, S., Zumer, J., Harris, J., Fine, E., Mehrotra, R., and Suleman, K. (2017). Frames: A corpus for adding memory to goal-oriented dialogue systems. *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219.

Bao, S., He, H., Wang, F., Wu, H., and Wang, H. (2020). PLATO: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96, Online, July. Association for Computational Linguistics.

Bentivogli, L., Cabrio, E., Dagan, I., Giampiccolo, D., Lo Leggio, M., and Magnini, B. (2010). Building textual entailment specialized data sets: a methodology for isolating linguistic phenomena relevant to inference. Valletta, Malta, May 2010.

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

Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., and 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*, pages 5016–5026, Brussels, Belgium, October-November. Association for Computational Linguistics.

Castellucci, G., Bellomaria, V., Favalli, A., and Romagnoli, R. (2019). Multi-lingual intent detection and slot filling in a joint bert-based model. *In ArXiv abs/1907.02884.*

Chakravarty, S., Chava, R. V. S. P., and Fox, E. (2019). Dialog acts classification for question-answer corpora. In *ASAIL@ICAIL.*

Colman, M., Eshghi, A., and Healey, P. (2008). Quantifying ellipsis in dialogue: an index of mutual understanding. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 96–99, Columbus, Ohio, June. Association for Computational Linguistics.

Cucurnia, D., Rozanov, N., Sucameli, I., Ciuffoletti, A., and Simi, M. (2021). Multi-annotator multi-language interactive light-weight dialogue annotator. *In EACL.*

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Eric, M., Goel, R., Paul, S., Sethi, A., Agarwal, S., Gao, S., Kumar, A., Goyal, A., Ku, P., and Hakkani-Tur, D. (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, pages 422–428, Marseille, France, May. European Language Resources Association.

Eshghi, A., Shalyminov, I., and Lemon, O. (2017). Interactional dynamics and the emergence of language games. *CEUR Workshop Proceedings*, 1863:17–21.

Fernández, R. and Ginzburg, J. (2002). Non-sentential utterances: A corpus-based study. *Traitement Automatique des Langues*, 43(2).

Han, T., Liu, X., Takanabu, R., Lian, Y., Huang, C., Wan, D., Peng, W., and Huang, M. (2021). Multiwoz 2.3: A multi-domain task-oriented dialogue dataset enhanced with annotation corrections and co-reference annotation. In Lu Wang, et al., editors, *Natural Language Processing and Chinese Computing*, pages 206–218, Cham. Springer International Publishing.

Hough, J. (2015). *Modelling Incremental Self-Repair Processing in Dialogue*. Ph.D. thesis, Queen Mary University of London.

Li, R., Kahou, S., Schulz, H., Michalski, V., Charlin, L., and Pal, C. (2018). Towards deep conversational recommendations. *In Advances in Neural Information Processing Systems 31 (NIPS 2018)*, pages 9748–9758.

Liu, X., Eshghi, A., Swietojanski, P., and Rieser, V., (2021). *Benchmarking Natural Language Understanding Services for Building Conversational Agents*, pages 165–183. Springer Singapore, Singapore.

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. *Proceedings of the SIGDIAL 2015 Conference*, pages 285–294.

Lowe, R., Pow, N., Serban, I. V., Liu, C.-W., and Pineau, J. (2017). Training end-to-end dialogue systems with the ubuntu dialogue corpus. *Dialogue and Discourse*, 8(1):31–65.

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 the Fourth International Conference on Language Resources and Evaluation (LREC'04).*

Mehri, S., Razumovskaia, E., Zhao, T., and Eskenazi, M. (2019). Pretraining methods for dialog context representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3836–3845, Florence, Italy, July. Association for Computational Linguistics.

Mohamad Suhaili, S., Salim, N., and Jambli, M. N. (2021). Service chatbots: A systematic review. *Expert Systems with Applications*, 184:115461.

Noble, B. and Maraev, V. (2021). Large-scale text pretraining helps with dialogue act recognition, but not without fine-tuning. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*. Association for Computational Linguistics.

Polignano, M., Basile, P., de Gemmis, M., Semeraro, G., and Basile, V. (2019). AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, volume 2481. CEUR.

Purver, M., Howes, C., Gregoromichelaki, E., and Healey, P. G. T. (2009). Split utterances in dialogue: A corpus study. In *Proceedings of the 10th Annual SIGDIAL Meeting on Discourse and Dialogue (SIGDIAL 2009 Conference)*, pages 262–271, London, UK, September. Association for Computational Linguistics.

Ramshaw, L. and Marcus, M. (1995). Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Ribeiro, E., Ribeiro, R., and de Matos, D. M. (2019). Deep dialog act recognition using multiple token, segment, and context information representations. *J. Artif. Intell. Res.*, 66:861–899.

Ritter, A., Cherry, C., and Dolan, D. (2010). Unsupervised modeling of twitter conversations. *In North American Chapter of the Association for Computational Linguistics (NAACL 2010).*

Schweter, S. (2020). Italian bert and electra models, nov.

Shriberg, E. (1996). Disfluencies in switchboard. In *In Proceedings of the International Conference on Spoken Language Processing*, volume 96, pages 3–6. Citeseer.

Sucameli, I., Lenci, A., Magnini, B., Simi, M., and Speranza, M. (2020). Becoming jilda. *In Proceedings of the Seventh Italian Conference on Computational Linguistics CLIC-it 2020.*

Sucameli, I., Lenci, A., Magnini, B., Simi, M., and Speranza, M. (2021). Toward data-driven collaborative dialogue systems: The jilda dataset. *IJCOL.*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, u., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Williams, J., Raux, A., Ramachandran, D., and Black, A. (2013). The dialog state tracking challenge. *In Special Interest Group on Discourse and Dialogue (SIGDIAL).*

Zhu, Q., Zhang, Z., Fang, Y., Li, X., Takanobu, R., Li, J., Peng, B., Gao, J., Zhu, X., and Huang, M. (2020). ConvLab-2: An open-source toolkit for building, evaluating, and diagnosing dialogue systems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 142–149, Online, July. Association for Computational Linguistics.

Appendix

Appendix A

Extract from a JILDA dialogue. The reported conversation is characterised by the presence of linguistic phenomena, such as proactivity and grounding (respectively, turn 1 and turns 7 & 8), and by sentences which are syntactically complex and semantically rich.

1. Applicant: *In the meantime, should I specify my geographic preferences? They could help you target me better*
(Nel frattempo potrei specificarti le mie preferenze a livello geografico? Potrebbero aiutarti nel targetizzarmi meglio)

2. Navigator: *Yes, perfect! Thank you*
(Sì, perfetto! Grazie)

3. 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*
(Attualmente vivo in Toscana: sono disponibile a trasferirmi in altre regioni ma anche all'estero non ho problemi di mobilità o limiti da questo punto di vista)

4. 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. I think they could be ok for you because both jobs involve management assignments and your degree could be very useful in these contexts.*
(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.)

5. Navigator: *Does one of them seem more appealing and do you want me to describe it first?*
(Uno dei due ti sembra più interessante e vuoi che te lo descriva per primo?)

6. 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?*
(Devo essere sincera: il primo non penso che possa fare al caso mio. Potresti descrivermi il secondo lavoro?)

7. Applicant: *I can't quite understand what "junior director" means*
(Non riesco a capire bene che cosa significhi "allievo direttore")

8. 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.*
(Certo! Le principali mansioni legate a questo impiego riguardano la pianificazione del budget e del conto economico dell'azienda. Il settore è quello alimentare quindi si tratta di compilare ordini e derrate alimentari, oltre che garantire la sicurezza sul lavoro e quella alimentare.)

Appendix B

We tried 12 different hyperparameter combinations on the validation set: three batch size values (32, 64, 128) and four learning rates ($1e-4$, $2e-5$, $3e-4$, and $5e-5$). Moreover, we kept the number of steps low to prevent overfitting, with check-step: 300 and max-step: 3000. The other relevant settings include *finetune*, *context* and *context-grad*. The fist one determines if the model will be tuned or not with the BERT parameter. If *fine-tune:false*, only added classification layers will be tuned.
The context parameter defines if use context information. If context: false, the [CLS] representation of the single utterance is passed to the intent classifier while the tokens' representations are passed to the slot classifier. If true, context utterances of the last three turns are concatenated and provide context information with embedding of [CLS] for dialogue act and slot classification.
Finally, context-grad determines whether compute the gradient through context representation, and then back-propagate the loss to the context encoder.
According to the results obtained evaluating the results on the validation set, we fixed the hyper-parameters as follows:

```
"model": {
"finetune": true,
"context": true,
"context_grad": false,
"check_step":300,
"max_step":3000,
"batch_size": 64,
"learning_rate": 1e-4,
"adam_epsilon": 1e-8,
"warmup_steps": 0,
"weight_decay": 0.0,
"dropout": 0.1,
"hidden_units": 768 }
```