From Cambridge to Pisa: A Journey into Cross-Lingual Dialogue Domain Adaptation for Conversational Agents

Tiziano Labruna[1,2], **Bernardo Magnini**[1]
1. Fondazione Bruno Kessler, Italy
2. Free University of Bozen-Bolzano, Italy
tlabruna@fbk.eu, magnini@fbk.eu

Abstract

English. Domain and language shift are still major bottlenecks for a vast range of task-oriented dialogue systems. This paper focuses on data-driven models for dialogue state tracking, and builds on top of recent work on *dialogue domain adaptation*, showing that state-of-the-art models are very sensible to language shift obtained through automatic translation. Experiments show that combining training data for the two languages (English and Italian) is always beneficial, while combining domains does not increase performance. As a relevant side effect of our work, we present a new dataset for dialogue state tracking available for Italian, derived from MultiWOZ 2.3.

Italiano. *I cambiamenti di dominio e di lingua sono ancora uno dei maggiori ostacoli per una ampia classe di sistemi di dialogo task-oriented. Questo lavoro si focalizza su modelli derivati da dati per tracciare gli stati del dialogo, e prosegue lavori recenti su adattamento del dialogo al dominio, mostrando che i modelli allo stato dell'arte sono molto sensibili ai cambiamenti di lingua ottenuti tramite traduzione automatica. Gli esperimenti mostrano che combinando i dati di addestranmento per due lingue (inglese e italiano) e' sempre vantaggioso, mentre la combinazione di dominii non migliora le prestazioni. Come importante conseguenza del lavoro, presentiamo il primo dataset per il tracciamento degli stati del dialogo disponile per l'italiano, derivato da MultiWOZ 2.3.*

Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-3033/paper71.pdf

1 Introduction

This paper is mainly motivated by the interest of exploring, and improving, the capacity of current data-driven task-oriented conversational systems to address shifts of domain and changes of language. Our starting point is the *dialogue domain adaptation* (DDA) approach proposed by (Labruna and Magnini, 2021), which allows to adapt training dialogues collected for a source domain knowledge (e.g., restaurants in Cambridge) to a domain where certain changes (e.g., a new restaurant opens, a restaurants changes its food , etc.) have occurred. The idea behind DDA is, that, rather then trying to improve the model robustness, it is worth to generate new training dialogues that are consistent with the domain changes. In this paper we extend and experiment DDA, so that also changes of language are included, particularly moving from English to Italian.

A further motivation for our work is related to the scarcity of annotated data for task-oriented dialogues systems for the Italian language. Among the dialogic resources originally collected in Italian, we mention the recent JILDA dataset (Sucameli et al., 2020) (Sucameli et al., to appear), a corpus of dialogues in the domain of job offers, semantically annotated. NESPOLE (Mana et al., 2004) is a corpus of human-human dialogues, which is not annotated for training purposes. The LUNA dataset (Dinarelli et al., 2010), although annotated with frame information, is not publicly available. Finally, the SUGAR dataset was presented at Evalita 2018 (Di Maro et al., 2018) allowing to train systems able to interpret spoken commands in Italian for the preparation of recipes.

A second type of data is derived through automatic translation from English. This is the case of the Italian version of the SNIPS dataset (Bellomaria et al., 2019), which provides annotations for spoken language understanding and can be

Dialogue Source	Dialogue Target
USER: I am looking for a **european** food restaurant in the expensive price range. Can you help with that?	USER: Sto cercando un ristorante con cibo **osteria** nella fascia di prezzo costosa. Puoi aiutarmi?
SYS: There are **5** of those. What area do you want to dine in?	SYS: Ce ne sono **7** di questi. In quale area vorresti cenare?
USER: In the centre of town please.	USER: In centro città per favore.
SYS: How about **eraina**? Shall I book you a table?	SYS: Che ne dici di **Hosteria Il Malandrone**? Ti prenoto un tavolo?
USER: Yes, please. It will be just me and I 'd like to eat at 21:00 on the same day as my train.	USER: Sì, per favore. Sarò solo io e vorrei mangiare alle 21:00 stesso giorno del mio treno.
SYS: OK, I 've got you booked. The reference number is **VMNDMKV2** and they 'll hold you table for 15 minutes.	SYS: OK, te l'ho prenotato. Il numero di riferimento è **WPQHRNE4** ti riserveranno il tavolo per 15 minuti.

Figure 1: Example of dialogue domain and language adaptation. Words in bold indicate slot-values that have been adapted.

used for slot filling and intent detection. Similarly, an Italian version of the SQuAD question-answer pair corpus is available (Croce et al., 2019) for Italian through translation. It has to be remarked that the Italian translations of both SNIPS and SQuAD do not rely on task-oriented dialogues, rather just on question-answer pairs. In addition, they are produced through simple translation from English, i.e., preserving domain information (e.g., names of places, restaurants, hotels, etc. reflect the English domain). We also notice that, unfortunately, the recent MultiATIS++ dataset (Xu et al., 2020), does not include Italian among the languages.

We are particularly interested in data-driven approach of dialogue state tracking (DST) (Balaraman and Magnini, 2021) for the Italian language. DST captures the capacity of a model to predict the correct *dialogue state* at each turn in a dialogue, representing both the communicative goals (dialogue acts) of the user and the portion of domain knowledge involved in such goals (slot-value pairs). To the best of our knowledge, the only dataset in Italian that can be used to model dialogue state tracking is JILDA (Sucameli et al., to appear), where dialogue state annotations were carried on following the MultiWOZ style. However, being concluded very recently, still there are no available DST baselines for JILDA, and, for this reason, we have developed an Italian version of the MultiWOZ dataset (Han et al., 2020).

Starting from MultiWOZ 2.3, a popular dataset in English developed for booking traveling facilities (e.g., restaurants, hotels, trains, attractions) in the area of Cambridge, we incrementally operated both language and domain shifts. We provide three experimental configurations: (i) a translation of the Cambridge data set into Italian; (ii) a domain shift from Cambridge to Pisa, maintaining English as language; and, finally, (3) a configuration where both the initial domain and the language are changed. As a relevant side effect, the datasets for the three configurations are now available for further research on dialogue state tracking for Italian[1].

In the paper we first introduce the relavant background in *dialogue domain adaptation* (Section 2), then we explain how dialogue domain adaptation is concretely applied to domain changes, and finally we report the experiments we have conducted (Section 4 and 5).

2 Dialogue Domain Adaptation

In the *Dialogue Domain Adaptation* setting (Labruna and Magnini, 2021), we assume an initial conversational domain, represented in a KB-SOURCE, and corresponding annotated training dialogues D-SOURCE. Then, as in real application scenarios, we assume that a number of changes occur in KB-SOURCE, such that a new conversational domain KB-TARGET needs to be considered. *Dialogue domain adaptation* consists in the capacity to automatically produce new annotated dialogues D-TARGET, such that they maintain both the linguistic structure and the linguistic variability of the initial D-SOURCE dialogues, while, at the same time, being consistent with the

[1] https://github.com/tLabruna/DDA

new KB-TARGET.

Figure 1 shows an example of dialogue adaptation. On the left side we have a user-system dialogue in English grounded on the Cambridge domain, while on the right side we have the same dialogue translated into Italian and adapted to the Pisa domain. In this paper we show how to generate such adapted dialogues (i.e. D-TARGET), which differ from the original dialogues (D-SOURCE) both in language and domain. The goal is then to train a dialogue state tracking model either on D-SOURCE or D-TARGET, and to investigate the impact of such adaptations on the model performance.

2.1 Slot-Value Substitution

Following (Labruna and Magnini, 2021), we focus on domain changes due to different slot-values, while assuming the same slot-names for both the source and target domains. As for language shift, it is based on translating all the utterances in a dialogue with the exclusion of the slot-values.

Given a slot-value occurring in a source dialogue D-SOURCE, the dialogue domain adaptation process consists of choosing the best slot-value in KB-TARGET to substitute the slot-value in the D-SOURCE utterance. The first step is to check whether the slot-value is known in KB-SOURCE. If it is known, we look for a correspondence in KB-TARGET, otherwise we directly keep it in D-TARGET (or, in case of different languages, translate it into target language). In order to decide if the slot-value is in the KB-TARGET, we use a similarity function based on a variation of the Gestalt Pattern Matching algorithm (Black, 2004). We select the most similar value in the KB-TARGET and we compare it to a empirically estimated threshold. Once we found a specific slot-value in KB-SOURCE and we ensured it exceeds the threshold, the corresponding slot-value to be selected from the KB-TARGET depends on the adaptation strategy we choose to adopt.

For the experiments of this paper we have used FREQUENCY-KB, an adaptation strategy based that obtained the best performance in (Labruna and Magnini, 2021). Given a slot-value in KB-SOURCE, FREQUENCY-KB basically consists of selecting the slot-value in KB-TARGET that has the most similar frequency distribution in the KB.

3 Method

We broke down the problem of adapting a conversational dataset to a new language and a new domain into three different steps: first we performed delexicalization by inserting some placeholders in the place of the slot values; then we automatically translated the dataset, leaving the placeholders unchanged; finally, we substituted the placeholders with the new domain slot-values. Each one of these steps is discussed in the following sub-sections.

3.1 Delexicalization

The setting that we are presenting involves the annotations being specifically slot-name slot-value pairs. Both the slot-values contained in the utterances, and those in the annotations, can not be translated the same way as the rest of the text, but need to undergo a Domain Adaptation process (e.g., we don't want *I need a taxi to The Old Castle* to be translated into *Ho bisogno di un taxi per Il Vecchio Castello*).

For this reason, the first step is to *delexicalize* a D-SOURCE dialogue, i.e., substituting all the slot-values in the utterances with placeholders. The example above shows this placeholder insertion, for moving from the following original sentence:

> "I need a restaurant in the north that has Caribbean food and a moderate price range please ."

to the utterance:

> "I need a restaurant in <#0#> that has <#1#> food and a <#2#> price range please ."

3.2 Translation

The second step is to perform the translation from the source language to the target language without considering the placeholders. According to our example, we will produce the following Italian utterance:

> "Ho bisogno di un ristorante a <#0#> che abbia <#1#> cibo e un <#2#> fascia di prezzo per favore ."

3.3 Slot-Value Substitution

As a third step, the placeholders need to be substituted back with slot-values of the target domain

KB-TARGET. Which slot-values to substitute depends on the Dialogue Domain Adaptation strategy and will be discussed later.

Finally, all the slot-values - both from utterances and annotations - that could not be substituted through DDA, need to be automatically translated, which will result in the following:

> "Ho bisogno di un ristorante a est che abbia caraibico cibo e un economico fascia di prezzo per favore ."

As can be noted, a downside of using placeholders is that this method does not consider the subject-verb agreement, nor the order of the words to be different between the original and the translated text. It should also be observed that in the cases of *north* and *moderate*, the slot substitution selects different values from the KB, while in the case of *Caribbean* it could not find a correspondence in the KB, hence it got translated directly from the original.

4 Experimental Setting

We started from the public available dataset MultiWOZ 2.3 (Han et al., 2020), which consists of a collection of more than ten thousand annotated dialogues (with dialogue states) spanning over seven domains related to traveling in Cambridge (e.g., restaurants, hotels, attractions, trains).

Pisa KB-TARGET. We manually created a KB-TARGET for Pisa, mirroring the instance distribution of the KB-SOURCE for Cambridge. For every entity instance of the Cambridge KB, a corresponding Pisa instance was created, keeping the slot-names as they were in the original, and changing only the slot-values. The specific instances were chosen by analysing the frequency distribution in the Cambridge KB and finding a similar correlation in the Pisa domain. For example, all the Cambridge restaurants with INDIAN food type, which is the most common in Cambridge, were substituted with Pisa restaurants with ITALIAN food type, which is the most common in Pisa. All the Pisa instances were taken from publicly available datasets containing real information on Pisa entities [2].

Automatic translation. As for translation from English to Italian, we used the automatic translator available at FBK. [3] The MT engine is built on the ModernMT framework[4] which features neural machine translation implementing the Transformer architecture (Vaswani et al., 2017). A big model (more than 200 million parameters) is trained on generic domain data, taken from the OPUS repository[5].

Test data used in the experiments were manually checked, correcting a number of translation issues, including, for instance, wrong prepositions used for time expressions (from *di 13:00* to *delle 13:00*), and wrong agreements (from *prezzi medio* to *prezzi medi*). Training data were not corrected.

Datasets. We run experiments over the following four datasets:

- CAM-ENG. This is the original MultiWOZ 2.3 dataset, with Cambridge as domain and English as language. It is used as referent for the other experiments.
- CAM-ITA. This is the translation to Italian of the original MultiWOZ 2.3 dataset, with Cambridge as domain.
- PISA-ENG. This is the original MultiWOZ 2.3 dataset adapted to the new Pisa knowledge base, using dialogue domain adaptation, as described in Section 3.
- PISA-ITA. This is the MultiWOZ 2.3 dataset, first translated into Italian and then domain adapted to the Pisa knowledge base.

For all the datasets we kept the same training/test split of dialogues as in the original MultiWOZ 2.3. In addition, we have experimented the following combinations:

- CAM-ITA + CAM-ENG. This combination provides all the available data for the Cambridge domain, mixing the two languages.
- PISA-ENG + CAM-ENG. This combination provides all the available data for English, mixing the two domains.
- CAM-ITA + PISA-ITA. This combination provides all the available data for Italian, mixing the two domains.

[2] http://www.datiopen.it/

[3] We would like to thank the Machine Translation Research Unit of FBK, and in particular Mauro Cettolo, for the kind support in the generation of automatic translations.

[4] http://github.com/modernmt/modernmt

[5] http://opus.nlpl.eu

Training	Test	Training Accuracy	Turn Accuracy	Joint F1	Joint Accuracy
Cam-ENG	Cam-ENG	0.52	0.97	0.9	0.49
Cam-ITA + Cam-ENG	Cam-ENG	0.48	0.97	0.9	0.49
Pisa-ENG + Cam-ENG	Cam-ENG	0.54	0.97	0.9	0.49
Cam-ITA	Cam-ITA	0.42	0.95	0.87	0.4
Cam-ITA + Cam-ENG	Cam-ITA	0.48	0.96	0.88	0.42
Cam-ITA + Pisa-ITA	Cam-ITA	0.4	0.95	0.87	0.38
Pisa-ENG	Pisa-ENG	0.54	0.97	0.89	0.5
Pisa-ITA + Pisa-ENG	Pisa-ENG	0.49	0.97	0.91	0.52
Pisa-ENG + Cam-ENG	Pisa-ENG	0.54	0.97	0.91	0.52
Pisa-ITA	Pisa-ITA	0.39	0.95	0.86	0.37
Pisa-ITA + Pisa-ENG	Pisa-ITA	0.49	0.96	0.88	0.42
Cam-ITA + Pisa-ITA	Pisa-ITA	0.4	0.95	0.86	0.37

Table 1: Performance of the TRADE algorithm over the datasets used in the experiments.

- PISA-ITA + PISA-ENG. This combination provides all the available data for the Pisa domain, mixing the two languages.

Dialogue State Tracking Model. The goal of the experiments is to assess the robustness of a dialogue state tracking model when domain and language are changed. As for DST model, we have used TRADE (Wu et al., 2019), an algorithm optimized for being used on multi-domain dialogues such MultiWOZ.

5 Results

Results of the experiments are presented in Table 1. The first column indicates which dataset the model was trained on; the second column reports the dataset used for testing the model; the last four columns report measures on the model performance. Training Accuracy refers to the Joint Accuracy obtained at training time; Turn Accuracy indicates how many single predictions were actually correct; the Joint F1 score reflects the accuracy of the model, considering both precision and recall; finally, the Joint Accuracy, measures the percentage of correct predictions of dialogue states for every dialogue turn, where a prediction is considered correct if all the slot values in the dialogue turn are correctly predicted. Results are reported into four groups depending on the dataset that has been used for testing. For every group we have three configurations: the first experiment reports the performance with the initial dataset, the second considers the extension of the initial dataset with the second language, and finally, the third experiment considers the extension of the initial dataset with the second domain.

6 Discussion

Results reveal several interesting aspects. First, we register a decrease in performance between the datasets in English and those automatically translated to Italian. This can be due to the process of placeholder insertion and subsequent substitution of slot-values, along with the translation itself, which can be source of errors. On the other side, the domain adaptation from CAM-ENG to PISA-ENG and from CAM-ITA to PISA-ITA did not show the same decrease of performance, rather it resulted even in a small increase for the first case.

The central part of our work, however, focused on generating adapted dialogues and investigating the performance variations derived from them.

Slot-name	Cam-ITA Accuracy	Cam-ITA vs Pisa-ITA Overlap	Cam-ITA + Pisa-ITA Accuracy	Cam-ITA vs Cam-ENG Overlap	Cam-ITA + Cam-ENG Accuracy
Train-departure	0.925	0.421	0.924	0.607	0.934
Train-destination	0.950	0.466	0.947	0.762	0.956
Restaurant-area	0.846	0.561	0.892	0.051	0.851
Hotel-area	0.787	0.812	0.811	0.03	0.795

Table 2: Slot-name accuracy prediction with comparison to the overlap of the slot-name between the dialogues. The first column is the considered slot-name. The second column is the predicted accuracy of the slot given by the TRADE model trained on Cam-ITA and tested on Cam-ITA. The third and fourth columns show the overlap and the prediction accuracy with respect to domain change. The remaining columns show the same measures for the language change.

With regards to this aspect, it should be noted that the addition of a second language resulted in a significant improvement almost in all cases, with an increase of 5% for CAM-ITA, 4% for PISA-ENG and 13.5% for CAM-ITA. On the other side, the addition of the second domain does not bring much advantage, resulting in zero change for CAM-ENG and PISA-ITA, a small decrease for CAM-ITA and a small increase for PISA-ENG.

6.1 Overlaps Between Datasets

In order to better understand the factors that affect the variation of Joint Accuracy performances between the datasets of each group, we have analysed the overlaps among the training datasets. We estimated such overlap as the proportion of slot-values in two datasets for every domain that are exactly the same .

We have observed that in most of the cases adding a dataset with high overlap for a certain domain produces an improvement in DST performance for that domain. As an example, the domain with highest overlap between the Cam-ITA dataset and the Pisa-ITA dataset is *Taxi* (86.11% of overlap). On the other side, the domain with lowest overlap between the same datasets is *Attraction* (44.45% of overlap). These overlaps have strong correlation with the DST performances on the two domains: the Cam-ITA + Cam-ENG dataset produces an improvement of 1.5 points with respect to the Cam-ITA dataset on the *Taxi* domain, and shows a decrease of 1 point on the *Attraction* domain.

This correlation can also be verified if we look at a slot-name level. Table 2 shows some examples of slot-names with corresponding overlaps between dialogues and slot-name prediction accuracy, taken from the Cam-ITA setting with domain and language additions. As it can be noted, when the slot-name overlap between the aggregated dialogue and Cam-ITA is higher, the respective prediction accuracy also tends to be higher.

7 Conclusion

We have investigated domain and language shift for data-driven task-oriented dialogue systems. We have extended recent work on *dialogue domain adaptation* to a cross-language setting, where both the domain and the language are changed. We showed that: (i) state-of-the-art models are very sensible to language shift obtained through automatic translation; (ii) combining training data for the two languages is always beneficial; on the contrary, combining data of different domains does not produce any improvement in all of our settings. Finally, as a relevant side effect of our work, we present a new dataset for dialogue state tracking available for Italian, derived from MultiWOZ 2.3. All the data are made available for further research on dialogue domain adaptation.

References

V. Balaraman and B. Magnini. 2021. Domain-aware dialogue state tracker for multi-domain dia-

logue systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:866–873.

Valentina Bellomaria, Giuseppe Castellucci, Andrea Favalli, and Raniero Romagnoli. 2019. Almawave-slu: A new dataset for SLU in italian. In Raffaella Bernardi, Roberto Navigli, and Giovanni Semeraro, editors, *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, volume 2481 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Paul E Black. 2004. Ratcliff/obershelp pattern recognition. *Dictionary of algorithms and data structures*, 17.

Danilo Croce, Alexandra Zelenanska, and Roberto Basili. 2019. Enabling deep learning for large scale question answering in italian. *Intelligenza Artificiale*, 13(1):49–61.

Maria Di Maro, Antonio Origlia, and Francesco Cutugno, 2018. *Overview of the EVALITA 2018 Spoken Utterances Guiding Chef's Assistant Robots (SUGAR) Task*, pages 79–85. 01.

Marco Dinarelli, Evgeny Stepanov, S. Varges, and Giuseppe Riccardi. 2010. The luna spoken dialogue system: Beyond utterance classification. pages 5366 – 5369, 04.

Ting Han, Ximing Liu, Ryuichi Takanobu, Yixin Lian, Chongxuan Huang, Wei Peng, and Minlie Huang. 2020. Multiwoz 2.3: A multi-domain task-oriented dataset enhanced with annotation corrections and co-reference annotation. *arXiv preprint arXiv:2010.05594*.

Tiziano Labruna and Bernardo Magnini. 2021. Addressing slot-value changes in task-oriented dialogue systems through dialogue domain adaptation. In *Proceedings of RANLP 2021*.

Nadia Mana, 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).

Irene Sucameli, 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, Bologna, Italy, March 1-3, 2021*, volume 2769 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Irene Sucameli, Alessandro Lenci, Bernardo Magnini, Manuela Speranza, and Maria Simi. to appear. Toward data-driven collaborative dialogue systems: The jilda dataset. *Italian Journal of Computational Linguistics*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Chien-Sheng Wu, 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.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual NLU. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5052–5063, Online, November. Association for Computational Linguistics.