Cross-Language Transformer Adaptation for Frequently Asked Questions

Luca Di Liello[‡], **Daniele Bonadiman**[‡][*]
Cristina Giannone[†], **Andrea Favalli**[†], **Raniero Romagnoli**[†], **Alessandro Moschitti**[‡]
[‡]DISI, University of Trento, Italy
[†]Almawave Srl., Italy
{luca.diliello,d.bonadiman,alessandro.moschitti}@unitn.it
{c.giannone,a.favalli,r.romagnoli}@almawave.it

Abstract

Transfer learning has been proven to be effective, especially when data for the target domain/task is scarce. Sometimes data for a similar task is only available in another language because it may be very specific. In this paper, we explore the use of machine-translated data to transfer models on a related domain. Specifically, we transfer models from the question duplication task (QDT) to similar FAQ selection tasks. The source domain is the well-known English Quora dataset, while the target domain is a collection of small Italian datasets for real case scenarios consisting of FAQ groups retrieved by pivoting on common answers. Our results show great improvements in the zero-shot learning setting and modest improvements using the standard transfer approach for direct in-domain adaptation [1].

1 Introduction

Frequently Asked Question (FAQ) websites are an essential service for user's self-assistance. FAQ websites typically present a list of questions, each associated with an answer. When searching for information, users have to go through the FAQs to determine whether there is a similar question providing a solution to their problem. However, this process does not scale well when the number of FAQs increases since too many questions may be presented to the user, and a simple search by the query may not retrieve the desired results. Additionally, in the last decade, users started looking for information using smartphones and voice assistants, such as Alexa, Google Assistant, or Siri. By design, voice assistants provide users with a different information access paradigm: the FAQ websites' navigation service is substituted by natural language dialogues, which satisfy the users' information need in few interactions. To achieve this goal, FAQ retrieval systems need to understand the question and present the user only with a set of strong candidates. One possible solution offered by personal assistants is constituted by (i) a FAQ retrieval system (Caputo et al., 2016) for efficiently finding relevant questions, and (ii) accurate neural models to select the most probable FAQ.

One of the major obstacles for building such a system is the availability of training data for the selection model. FAQ systems are domain-specific in nature since they aim to provide users with information about specific websites or services. Moreover, the industrial setting does not always allow for creating a large corpus of questions for any specific domain, as the customers (FAQ's owners) typically cannot provide such data. There are many reasons: (i) they are not familiar with the process of training data creation, as it is not part of their business; (ii) the topic of the FAQ system does not require more than tens of question/solution pairs; (iii) it is not easy to generate a dataset for question-question similarity from a question-answer system.

A traditional approach to alleviating such a problem is to use transfer learning (TL), i.e., data from other domains/tasks is used to train a model on the target task. TL research has been boosted by the availability of pre-trained transformer-based models (Vaswani et al., 2017; Devlin et al., 2018), which capture general-purpose language models. In this paper, we approach the problem of FAQ selection, fine-tuning pre-trained language models on the Question Duplication Task (QDT) from Quora[2]. This task aims to identify whether

[*] work done prior to joining Amazon

[1]

[2] https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs

Task	Question 1	Question 2	Label
QDT	How many months does it take to gain knowledge in developing Android apps from scratch?	How much time does it take to learn Android app development from scratch?	True
QDT	How do I prepare for software interviews?	What are the best ways to prepare for software interviews?	True
QDT	Why did harry become a horcrux?	What is a Horcrux?	False
QDT	What is journalism for you?	What is journalism?	False
FAQ	Can medicines be sold on Amazon?	What items can't I sell on Amazon?	True
FAQ	I forgot my username	Why won't the page load?	True
FAQ	Is it possible to change my personal information after I have registered?	Is it possible to change the password?	False
FAQ	Can I have food brought from home during the flight?	What is included in the price I pay?	False

Table 1: Some examples of QDT and FAQ pairs. Notice that in the first block question are paraphrase of each other. The second block contains instead questions that only share a common answer.

two questions are duplicated or not, i.e., semantically equivalent or not. (Androutsopoulos and Malakasiotis, 2010).

Although the FAQ selection task shares some commonalities with QDT one, they are different. A FAQ task can indeed be solved by ranking all the FAQs in the collection using a system that computes the semantic similarity score between two questions, i.e., a Paraphrase Identification model. However, there are still some crucial differences. While QDT requires to infer if two questions are semantically equivalent, FAQ selection seeks questions that share the same intent and, at the same time, that they share the same answer. Moreover, the FAQ selection strongly depends on the domain in which the retrieval system is applied. For example, if a website responds to every technical complaint with "contact us", there will be many positive pairs that will not share any real answer. Every portal in which a FAQ similarity system is needed, e.g., online services and e-commerce, requires a different level of details depending on the service type and its complexity. Table 1 provides some examples taken from QDT and FAQ datasets to underline the difference better.

One of the largest corpora for the fine-tuning of QDT is the well-known Quora dataset, sourced from the homonymous community question answering website. The dataset is constituted by question pairs, labeled as being duplicates or not. However, the Quora dataset is only available in the English language, preventing its use for building Italian systems.

In this paper, we propose to adapt Transformer architectures to the task of FAQ selection using machine translation. We first translated the Quora dataset to Italian, and then we trained a state-of-the-art QDT model for Italian. Finally, we tested the adapted QDT model to two FAQ datasets showing significant improvement on the zero-shot learning baselines (i.e., using no target domain training data). Moreover, we show that fine-tuning the adapted model on small target data provides a consistent improvement over models not exploiting our transfer learning approach. It should be noted that our techniques can be seen as an extension of the Transfer and Adapt (TANDA) (Garg et al., 2019), but with the difference that transfer is carried out on a similar approximate task using translated data, i.e., Approximated machine Translated TANDA (ATTANDA).

The rest of the paper is organized as follows: Section 2 describes similar approaches to do Cross-Lingual Transfer Learning, Section 3 provides an overview of the available datasets and Section 4 describes the methodology we developed. Finally, Section 5 summarizes the main results and Section 6 draws the conclusions of this work.

2 Related Works

The current state of the art for QDT makes use of pre-trained transformer-based frameworks, e.g., BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019) or XLNet (Yang et al., 2019). These models have millions of parameters that are trained in a two-step approach. First, they are trained as language models using various losses (e.g., masked language modeling or sentence order prediction loss) on a large corpus in an unsupervised way and then are fine-tuned on the target labeled dataset.

In Transfer Learning, a model is *transferred* (i.e., trained) on data coming from a high-resource task and is then adapted to another, usually more specific. All the Transformers-based models can be seen as Transfer Learning models: they are first trained on large corpora of unlabelled data and then are specialized in a downstream domain. Nonetheless, there are scenarios where data about similar tasks can further improve already-great models.

Cross-lingual transfer-learning (CLTL) is an extension in which data from a high-resource language is used to solve a low-resource language task. This technique is sometimes used in combination with Cross-Lingual Word Embeddings alignment. The actual trend is to align word embeddings to focus only on shared language-independent features and then apply Transfer Learning techniques (Lange et al., 2020; Keung et al., 2020). However, solving a task using data coming from a similar one has different requirements.

A similar approach to our has been explored by (Schuster et al., 2019), in which they used multilingual data to improve the performance of low-resource languages. However, even if they used translated data, they did not explore applying the transferred model to an affine task. Another approach (Do and Gaspers, 2019) filters high-quality samples from a high-resource language dataset to train the model in reduced time. Authors claim a significant improvement in the target language and task, even using only a small amount of computing.

In (Joty et al., 2017), the authors improve the performance in question-question similarity by using an adversarial approach. Thanks to adversarial training, they extract language-independent features from a trained model with supervision on a high-resource language and adapted to a low-resource one for testing. Results show important improvements in the target language, even in the zero-shot setting.

Also, in (Wang et al., 2020), a complete overview of the common approaches for cross-lingual transfer learning (CLTL) is proposed. Authors start by comparing (i) *joint training*, in which a model is trained on multilingual data using both a monolingual and a cross-lingual loss, and (ii) *CLWE alignment* before training, in which language embeddings are mapped to a shared space before fine-tuning. They find out that both methods perform well and that there is not an overall winner. Finally, they show that training with both approaches outperforms previous state-of-the-art methods.

3 Datasets

3.1 Quora Question Pairs

The Quora dataset is a collection of question pairs for QDT. It contains many semantically equivalent questions that people asked more than once, for example, ”What is the most populous state in the USA?” and ”Which state in the United States has the most people?”. Human experts have assigned labels; therefore, it is not free from subjective decisions and questionable labels. The dataset contains about $404K$ question pairs, 37% with a positive label, and 63% with a negative one. However, this dataset is not error-free: many ids are used more than once ($14K$), and many questions are referred by more than a single id ($76K$).

3.2 FAQ: RDC and LCN

RDC and LCN are two real-world datasets of FAQ retrieval. They were designed to build a QA component of conversational agent systems in Italian, targeting specific domains. Neither dataset is ready for FAQ retrieval out of the box, so we needed to group questions differently. Given that many questions share a common answer in RDC, we created several examples for the FAQ selection task by clustering questions with respect to the answers. For RDC, since the answers were simply the name of the category in which an answer could be found, we pivoted on the categories to create the clusters.

To build the examples, we first built clusters of equivalent questions, using their similarity gold standard labels, or rather the answers or the categories. LCN consists of 388 questions, which we grouped in 24 clusters of different sizes. The smallest contains only two elements, while the largest contains 50 elements. RDC contains 369 entries, which we grouped in 30 clusters with a minimum and maximum size of 1 and 37, respectively. [3]

Tests will show that LCN is the hardest dataset. The reason is that clustering has not been applied by pivoting on the answers but the same category instead (answers were not available). Then, each cluster contains questions that do share a precise answer but rather the same category.

[3] There is an Italian FAQ dataset called QA4FAQ, but it is not suitable for question similarity since annotations for the dataset are not available.http://qa4faq.github.io

The transformation of a set of clusters in a training or test set was done with the following algorithm: for N times, an element from each cluster was chosen, called champion, and was temporarily removed from its cluster. Each champion was then paired with a random element from every cluster, assigning positive labels when the two shared belonging to the same cluster. We found that $N = 5$ was a reasonable number of rounds since more would have lead to information repetition.

Moreover, there was a need to create both small training and test sets to measure models' performance when fine-tuned on the FAQ domain. We could not divide the dataset described before since training and test sets would have had many common sentences. To accomplish a perfect separation, 70% of the clusters were used to create a train set while the remaining 30% were used for the test set.

3.3 FAQ: ItaFAQ

We built a small FAQ dataset in Italian by scraping popular websites. Then, we asked 10 different people with different backgrounds and levels of education to create additional questions similar to those automatically collected. The specific request was to create questions that would have had the same or a similar answer. The dataset is released as open-source and is available for download[4]. This dataset can be useful to test an information retrieval system. However, it is easier to solve than the previously described RDC and LCN. The main reasons are that (i) humans tend to create partially related new questions, and that (ii) general FAQ dataset about well-known companies and topics are easier to process than strong domain-specific data.

4 ATTANDA Approach

4.1 Machine Translation of Quora

There are no medium or large-size Italian datasets for QDT or FAQ retrieval; thus, we applied machine translation. We used *Microsoft Azure Cognitive Services* to translate Quora Question Pairs into Italian. Since the original Quora dataset had some questions repeated on different entries, we followed the approach in (Haponchyk et al., 2018; Bonadiman et al., 2019) and grouped all the questions in clusters by mean of the transitive property: if a and b are the two questions of a pair with a positive label and C_i is a cluster, $a \in C_i \leftrightarrow b \in C_i$. Moreover, if there is a tuple (a, b) with a positive label and $a \in C_i, b \in C_j$, then C_i and C_j are merged in $C_k = C_i \cup C_j$.

After that, we translated all the questions of the clusters with at least two members. This allowed us to effectively reduce machine translation costs because we avoided translating questions that would have appeared only in negative pairs (millions of negative pairs can be easily generated by randomly picking questions from different clusters). We built the transfer dataset by labeling (i) all pairs of questions in the same cluster as positive examples; and (ii) a random number of pairs with members from different clusters as negative examples. We limited the number of the latter to be equal to the number of positive examples.

4.2 Transformer architectures

To reach the highest performance, we developed our models on the actual state of the art for QA. We took into consideration:

- **Multilingual BERT** (mBERT), a BERT model trained on the 104 largest Wikipedia, in terms of the number of articles. The model contains 177M[5] parameters and has 12 transformer layers (Devlin et al., 2018);
- **Italian BERT**[6], a BERT model trained only on Italian text. The version we used was trained over the concatenation of the OSCAR corpus and the Italian OPUS corpus, for a total of 81GB of text. This model features a total of 110M parameters on 12 layers;
- **GilBERTo**[7], a RoBERTa model trained over 71GB of lowercase Italian text extracted from the OSCAR corpus. The authors state that this model applies masking to whole words (WWM), as in (Martin et al., 2020), instead of masking at the sub-words level, as in the original BERT. This model has a total of 111M parameters.

[4] The dataset can be downloaded at https://github.com/lucadiliello/italian-faq-dataset

[5] mBERT has a bigger size since its vocabulary is considerably larger than monolingual models.

[6] Italian BERT models and code are available at https://github.com/dbmdz/berts

[7] GilBERTo models and code are available at https://github.com/idb-ita/GilBERTo

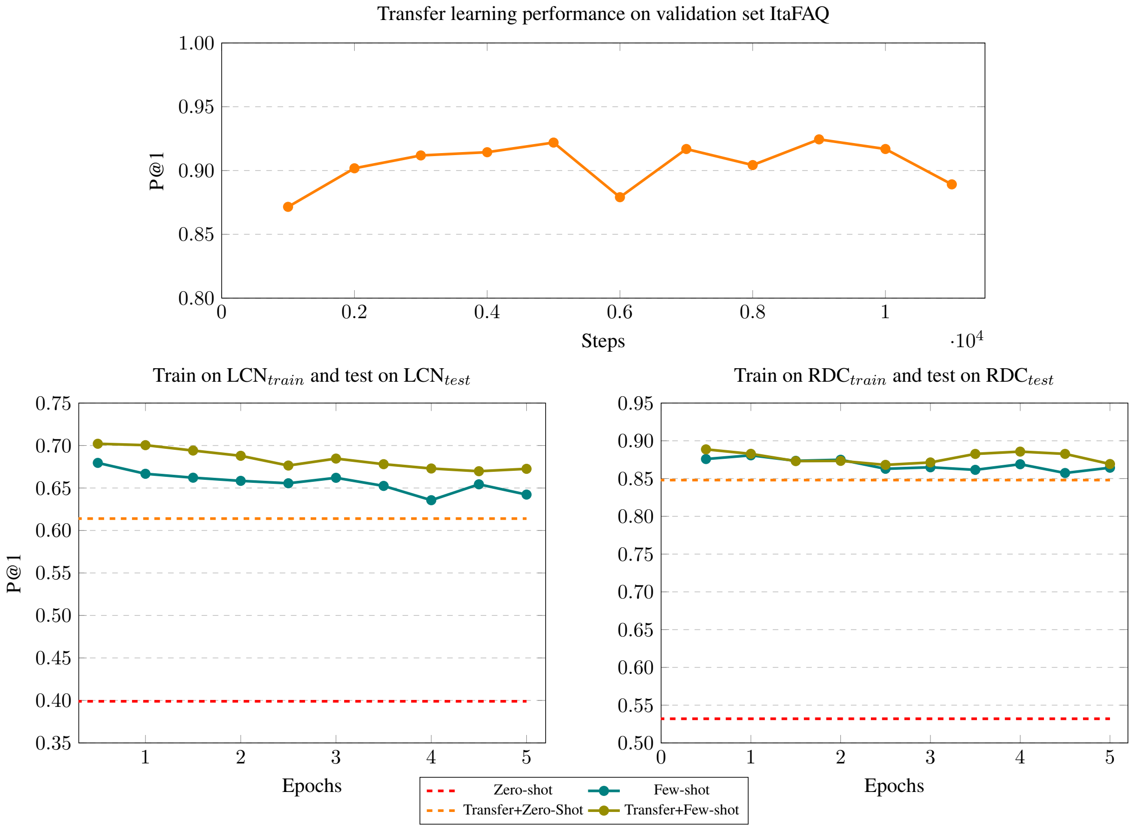

Figure 1: **Above**: Trend of the P@1 when transferring on Quora and validating on ItaFAQ. **Below**: Comparison of *Zero-Shot* (no training at all on Quora nor RDC$_{train}$ or LCN$_{train}$), *Few-Shot* learning (only in-domain adaptation on RDC$_{train}$ or LCN$_{train}$), *Transfer+Zero-Shot* learning (transfer on Quora and test directly) and *Transfer+Few-Shot* learning (transfer on Quora and in-domain adaptation on RDC$_{train}$ or LCN$_{train}$. The results are the average of 8 runs with different seed and dataset splits. All runs used the same hyper-parameters: batch size of 32 and Adam optimizer with a learning rate of $1e - 05$. Notice the different scales on the y-axis.

4.3 Cross-Domain training

We aim at exploiting data similar to the target task, which may also come from a different language, to train models for our FAQ target task. Our approach can be seen as an extension of TANDA by (Garg et al., 2019), which consists in two-step fine-tuning. First, they transfer the model on a general QA task with a huge dataset, and then they adapt the model to a smaller and specific QA benchmark such as WikiQA. They showed that a transfer step could improve the final performance if the source and target tasks are similar. We extend this idea by creating our transfer dataset utilizing machine translation, as described before. We call our approach ATTANDA (Approximated machine-Translated TANDA).

5 Results

This section shows the results of testing different models on the FAQ retrieval task. We use Precision at 1 (P@1), which is equal to accuracy, as we mainly need to measure if the returned FAQ is correct. LCN$_{train}$, LCN$_{test}$, RDC$_{train}$ and RDC$_{test}$ are the names of the splits of LCN and RDC derived by dividing the set of clusters.

We start by comparing the available, transformer-based models. Table 2 shows that Italian BERT is better than the other models in most tests. This comes not as a surprise since it is specialized in the Italian language, it takes into consideration the case sensitivity of the input text, and it is trained on the most extensive corpus. GilBERTo also performs well, but RoBERTa's improvement is insufficient to overcome the smaller training set and the case-insensitive tokenizer.

Once we established that the best pre-training model is Italian BERT, since it shows the highest scores in 3 comparisons out of 4, we tested different transfer methods on LCN and RDC splits. We compare the performance of Italian BERT in two scenarios: (i) the model is directly fine-tuned on the target domain, and (ii) the model is first transferred on Quora and then fine-tuned on the target domain (ATTANDA). We also report the results of the model without in-domain fine-tuning

Models	Dataset		Results	
	Train	Test	MRR	P@1
mBERT	-	LCN$_{test}$	45.4	25.5
IT BERT	-	LCN$_{test}$	**56.6**	**39.9**
GilBERTo	-	LCN$_{test}$	47.0	29.1
mBERT	-	RDC$_{test}$	59.4	41.0
IT BERT	-	RDC$_{test}$	65.1	53.2
GilBERTo	-	RDC$_{test}$	**67.9**	**56.1**
mBERT	Quora	LCN$_{test}$	64.3	49.2
IT BERT	Quora	LCN$_{test}$	**75.1**	**61.4**
GilBERTo	Quora	LCN$_{test}$	72.4	58.2
mBERT	Quora	RDC$_{test}$	88.4	81.1
IT BERT	Quora	RDC$_{test}$	**91.1**	**84.8**
GilBERTo	Quora	RDC$_{test}$	89.7	83.3

Table 2: Comparison of different transformers-based models. Each model in the bottom half of the table has been trained on Quora with the same hyper-parameters (batch size of 64 and Adam optimizer with a learning rate of $1e-05$) for a single epoch. Reported metrics are the average over 8 runs with different seeds and splits.

(zero-shot tests), taken from table 2.

Figure 1 reports the P@1 while training for the first five epochs on the test sets. This does not affect consistency of results since we do a comparison on the whole fine-tuning phase. All the plots show that transferring the model first on Quora gives an increase in P@1, especially in the early steps. Also, in this setting, training for more than two epochs did not provide further improvement, which could lead to over-fitting. This is intuitive as the training and test splits are small and also contain repeated information. There is no clear reason not to perform a transfer step since the resulting performance is at least equal and the computational effort to train for a single epoch on Quora is negligible.

6 Conclusion

We explored transfer learning in a typical industrial scenario where only small (or no) data is available in the target language. We showed that it is possible to use machine translated data to improve a strictly related task's performance. We suspect that if the tasks had been more similar, for example, Question Answering and FAQ, the performance gain would have been even better. However, this was a real-world scenario where the target datasets were used for production in real websites, and size and quality were not large. In this setting, applying a transfer phase can improve the retrieval of similar questions, and the transfer step is a low-cost operation compared to the pre-training.

References

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.

Daniele Bonadiman, Anjishnu Kumar, and Arpit Mittal. 2019. Large scale question paraphrase retrieval with smoothed deep metric learning. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 68–75.

Annalina Caputo, Marco de Gemmis, Pasquale Lops, Francesco Lovecchio, Vito Manzari, and Acquedotto Pugliese AQP Spa. 2016. Overview of the evalita 2016 question answering for frequently asked questions (qa4faq) task. In *of the Final Workshop 7 December 2016, Naples*, page 124.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Quynh Do and Judith Gaspers. 2019. Cross-lingual transfer learning with data selection for large-scale spoken language understanding. pages 1455–1460, 01.

Siddhant Garg, Thuy Vu, and Alessandro Moschitti. 2019. Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection.

Iryna Haponchyk, Antonio Uva, Seunghak Yu, Olga Uryupina, and Alessandro Moschitti. 2018. Supervised clustering of questions into intents for dialog system applications. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2310–2321, Brussels, Belgium, October-November. Association for Computational Linguistics.

Shafiq Joty, Preslav Nakov, Lluís Màrquez, and Israa Jaradat. 2017. Cross-language learning with adversarial neural networks. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 226–237, Vancouver, Canada, August. Association for Computational Linguistics.

Phillip Keung, Yichao Lu, and Vikas Bhardwaj. 2020. Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and ner.

Lukas Lange, Anastasiia Iurshina, Heike Adel, and Jannik Strötgen. 2020. Adversarial alignment of multilingual models for extracting temporal expressions from text.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv*, pages arXiv–1907.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime Carbonell. 2020. Cross-lingual alignment vs joint training: A comparative study and a simple unified framework.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.