Deep Bidirectional Transformers for Italian Question Answering

Danilo Croce and **Giorgio Brandi** and **Roberto Basili**
Department Of Enterprise Engineering
University of Roma, Tor Vergata
Via del Politecnico 1, 00133 Roma
{croce,basili}@info.uniroma2.it *

Abstract

English. Deep learning continues to achieve state-of-the-art results in several NLP tasks, such as Question Answering (QA). Unfortunately, the requirements of neural QA systems are very strict in the size of the involved training datasets. Recent works show that the application of Automatic Machine Translation is an enabling factor for the acquisition of large scale QA training sets in resource poor languages such as Italian. In this work, we show how these resources can be used to train a state-of-the-art deep architecture, based on effective techniques recently proposed within the Bidirectional Encoder Representations from Transformers (BERT) paradigm.

Italiano. *I recenti studi sull'applicazione di metodi di Deep Learning hanno portato a risultati importanti rispetto a diversi problemi di Natural Language Processing, come il Question Answering (QA) task. Sfortunatamente, i requisiti di tali sistemi di QA neurali sono molto stringenti per quanto riguarda le dimensioni dei dataset necessari per addestrare i modelli piú complessi. Tuttavia, recenti lavori hanno dimostrato che é possibile applicare tecniche di traduzione automatica al fine di acquisire collezioni di esempi di larga scala e addestrare architetture neurali per il Question Answering nelle lingue in cui i dati di training sono scarsi, come l'italiano. In questo lavoro, mostriamo come queste risorse permettono l'addestramento di una architettura neurale molto efficace, basata sul paradigma noto come Bidirectional Encoder Representations from Transformers (BERT), con risultati che costituiscono lo stato dell'arte.*

1 Introduction

Question Answering (QA) ((Hirschman and Gaizauskas, 2001)) tackles the problem of returning one or more answers to a question posed by a user in natural language, using as source a large knowledge base or, even more often, a large scale text collection: in this setting, the answers correspond to sentences (or their fragments) stored in the text collection. A typical QA process consists of three main steps: the question processing that aims at extracting requirements and objectives of the user's query, the retrieval phase where documents and sentences that include the answers are retrieved from the text collection and the answer extraction phase that locates the answer within the candidate sentences (Harabagiu et al., 2000; Kwok et al., 2001).

Various QA architectures have been proposed so far. Some of these rely on structured resources, such as Freebase, while others use unstructured information from sources such as Wikipedia (an example of such a system is the Microsoft's AskMSR (Brill et al., 2002)), or generic Web pages, e.g. the QuASE system (Sun et al., 2015). Hybrid models exist as well, that make use of both the structured and the unstructured information. These include IBM's DeepQA (Ferrucci et al., 2010) and YodaQA (Baudiš and Šedivý, 2015).

In order to initialize such systems, a manually constructed and annotated dataset is crucial, from which the mapping between questions and answers can be learned. Datasets designed for structured-knowledge based systems, such as WebQuestions (Berant et al., 2013), usually contain the questions, their logical forms and the answers. On the other side, datasets over unstructured information are usually composed of question-answer

pairs: WikiMovies (Miller et al., 2016) is an example of this class of systems and it is made of a collection of texts from the movie domain. Finally, some datasets contain the entire triplets made of the questions, the paragraphs and the answers, that are expressed as specific spans of the paragraph and thus located in the paragraph. This is the case of the recently proposed SQuAD dataset (Rajpurkar et al., 2016).

State-of-the-art approaches proposed in literature (Chen et al., 2017; Seo et al., 2017; Clark and Gardner, 2018; Peters et al., 2018) are based on neural paradigms and are often portable across different languages. Among them, the neural approach presented in (Devlin et al., 2019), beside achieving state-of-the-art results in several NLP tasks, is shown competitive in QA even with respect to human annotators.

Unfortunately, the limited availability of training data for languages different from English still remains an important problem. Even though multilingual data collections, such as Wikipedia, do exist for many languages, the portability of the corresponding annotated resources for supervised learning algorithms remains limited: large-scale annotated data mostly exist only for the English language.

Recent works show that the application of Automatic Machine Translation enables the acquisition of large corpora for QA in resource poor languages such as Italian (Croce et al., 2018; Croce et al., 2019). As a result, SQuAD-IT, i.e., a large scale dataset made of about 50,000 questions/answer pairs has been made available. It was not fully manually validated but still represents a valuable resource for training neural approaches.

In this work, we show how these resources enable the training of a recent and promising deep neural architecture, based on the effective techniques recently justified within the Bidirectional Encoder Representations from Transformers (BERT) paradigm (Vaswani et al., 2017; Devlin et al., 2019). The experimental evaluation carried out with respect to SQuAD-IT confirm the impressive results of BERT even in Italian QA, providing state-of-the-art results which are far higher with respect to previous methods.

In the rest of the paper, section 2 introduces the BERT architecture for QA. Section 3 report the experimental evaluation, while Section 4 draws some conclusions.

2 Bidirectional Encoder Representations for QA

In the field of computer vision, researchers have repeatedly shown the beneficial contribution of transfer learning, i.e., the pre-training a neural network model on a known task, for instance image classification with respect to the ImageNet dataset, and then performing fine-tuning using the trained neural network as the basis of a new purpose-specific model, e.g., (Girshick et al., 2013).

The approach proposed in (Devlin et al., 2019), namely Bidirectional Encoder Representations from Transformers (BERT) provides a very effective model to pre-train a deep and complex neural network over very large scale of unannotated texts and to apply it to a large variety of NLP task by simply extending it to each new problem by fine-tuning the entire architecture.

The building block of BERT is the *Transformer* element, an attention-based mechanism that learns contextual relations between words (or sub-words, i.e. word pieces, (Schuster and Nakajima, 2012)) in a text. In its original form, proposed in (Vaswani et al., 2017), Transformer includes two separate mechanisms, an encoder that reads the text input and a decoder that produces a prediction for the targeted Machine Translation tasks.

In line with (Peters et al., 2018), BERT aims at providing a sentence embedding (as well as the contextualized embeddings of each word composing the sentence) where the pre-training stage aims at acquiring an expressive and robust language model, where only the encoder is used. As shown in Figure 1 (on the left) the Transformer encoder reads the entire sequence of words at once and acquire a language model by reconstructing the original sentence applying a MLM (*masked language model*) pre-training objective: the MLM randomly masks some of the tokens from the input, and the objective is to predict the original masked word based only on its context. In addition to the masked language model, BERT also uses a *next sentence prediction* task that jointly pre-trains text-pair representations. This last objective is crucial to improve the network capability of modeling relational information between text pairs, which is particularly important in tasks such as QA in order to relate an answer to a question.

After the language model is trained over a generic document collection, the BERT architecture allows encoding (i) specific words belong-

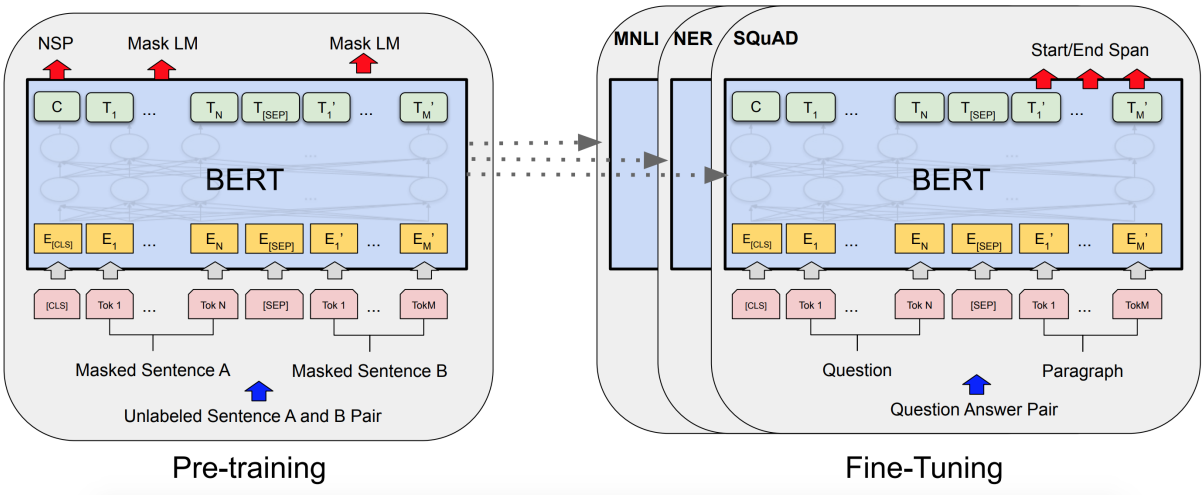

Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token (e.g. separating questions/answers).

ing to a sentence, (ii) the entire sentence and (iii) sentence pairs with dedicated embeddings. These can be used in input to further deep architectures to solve sentence classification, sequence labeling or relational learning tasks by simply adding simple layers and fine-tuning the entire architecture. On top of such embeddings, *fine-tuning* is applied by adding task specific and simple layers on top of the architecture acquiring the language model. In a nutshell, this layer introduces minimal task-specific parameters, and is trained on the targeted tasks by simply fine-tuning all pre-trained parameters, optimizing the performance on the specific problem. The straightforward application of BERT has shown better results than previous state-of-the-art models on a wide spectrum of natural language processing tasks.

One of the most impressive results was achieved with respect to the Question Answering task proposed by (Rajpurkar et al., 2016): given a question and a passage from Wikipedia containing the answer, the task is to predict the answer text span in the passage. An example of paragraph, showing the Wikipedia answer to the question "*What was Marie Curie the first female recipient of?*" is reported in Figure 2. This specific task originated the Stanford Question Answering Dataset (SQuAD), a collection of 100k crowd-sourced question/answer pairs.

The fine-tuning process of BERT in the QA task (shown on the right side of Figure 1) requires to encode the input question and passage as a generic text pair, such as the ones used for the next sentence prediction task used in the initial training stages.

In order to determine the correct span for the answer, (Devlin et al., 2019) introduces on top of embeddings encoding the words of the question/answer pairs a so-called *start vector* $S \in \mathcal{R}^H$ (with H the dimensionality of the embedding produced for each wordpiece T_i) and an *end vector* $S \in \mathcal{R}^H$. Then, the probability of word i being the start of the answer span is computed as a dot product between the associated embedding T_i and S followed by a softmax layer over all of the words in the paragraph: $P_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}}$. The analogous formula is used for the end of the answer span. The score of a candidate span from position i to position j is defined as $S \cdot T_i + E \cdot T_j$, and the maximum scoring span where $j \geq i$ is used as a prediction. The training objective is the sum of the log-likelihoods of the correct start and end positions. The above fine-tuning of BERT achieved state-of-the-art results over the official benchmarking campaign related to SQuAD and, most noticeably, its accuracy is comparable to the one observed in human annotators[1].

It is worth noting that no bias over the input lan-

[1] https://rajpurkar.github.io/SQuAD-explorer/

QUESTION: *What was Marie Curie the first female recipient of?*

WIKIPEDIA PARAGRAPH: One of the most famous people born in Warsaw was Maria Skłodowska-Curie, who achieved international recognition for her research on radioactivity and was the first female recipient of the **Nobel Prize.**[198] Famous musicians include Władysław Szpilman and Frederic Chopin. Though Chopin was born in the village of Żelazowa Wola, about 60 km (37 mi) from Warsaw, he moved to the city with his family when he was seven months old.[199] Casimir Pulaski, a Polish general and hero of the American Revolutionary War, was born here in 1745.[200]

GROUND TRUTH ANSWER: *Nobel Prize*

Figure 2: An example of the SQuAD dataset (Rajpurkar et al., 2016).

Element	Training set			Test set		
	English	Italian	Percent.	English	Italian	Percent.
Paragraphs	18,896	18,506	97.9%	2,067	2,010	97.2%
Questions	87,599	54,159	61.8%	10,570	7,609	72.0%
Answers	87,599	54,159	61.8%	34,726	21,489	61.9%

Table 1: The quantities of the elements of the final dataset obtained by translating the SQuAD dataset, with the percentage of material w.r.t the original dataset. The Italian test set was obtained from the English development set, being the English test set not available publicly.

	DrQA-IT	BERT-IT
EM	56.1	64.96
F1	65.9	75.95

Table 2: Results of BERT-iT over the SQuAD-IT dataset

guage exists, so that the language model underlying BERT can be acquired over any text collection independently from the input language. As a consequence a pre-trained model acquired over documents written in more than one hundred languages exists. It will be applied in the next section to train and evaluate such a QA model over a dataset of examples in Italian.

3 Experimental Evaluation

In order to assess the applicability of the BERT architecture against the targeted QA task, a multilingual pre-trained model has been downloaded[2]: in particular, this model has been acquired over documents written in one hundred languages, it is composed of 12 layers of Transformers and associates each token in input to a word embedding made of 768 dimensions. For consistency with (Devlin et al., 2019), 5 epochs have been considered to fine-tune the model.

We trained the architecture over SQuAD-IT[3], a dataset made available by (Croce et al., 2019). This dataset includes more than 50,000 question/paragraph pairs obtained by automatic translating the original SQuAD dataset. The details about the number of sentences is reported in Table 1 where a comparison with the original SQuAD in English is reported.

The parameters of the neural network were set equal to those of the original work, including the word embeddings resource. Two evaluation metrics are used: exact string match (EM) and the F1 score, which measures the weighted average of precision and recall at the token level. EM is a stricter measure evaluated as the percentage of answers perfectly retrieved by the systems, i.e. the text extracted by the span produced by the system is exactly the same as the gold-standard. The adopted token-based F1 score smooths this constraint by measuring the overlap (the number of shared tokens) between the provided answers and the gold standard.

Performances are reported in Table 2 together with the results achieved by a variant of the DrQA system (Chen et al., 2017), evaluated against the same SQuAD-IT dataset, as from (Croce et al., 2019). Improvements are impressive, as both EM and F1 are improved of more than 10%. Anyway, these results are in line with the impact of BERT over the original English dataset. In the final version of this paper we will provide an in depth comparison between DrQA and BERT.

[2] https://storage.googleapis.com/bert_models/2018_11_23/multi_cased_L-12_H-768_A-12.zip

[3] https://github.com/crux82/squad-it

4 Conclusions

This paper explores the application of Bidirectional Encoder Representations within the QA task in Italian, enabled by the recent availability of a large-scale annotated corpus, SQuAD-IT. The experimental results confirm the robustness of the adopted Transformer-based architecture, with a significant improvement with respect to earlier neural architectures. This result paves the way to the development of portable, robust and accurate neural models for QA in Italian, and future work will certainly consider other possible extensions of the adopted model.

References

Petr Baudiš and Jan Šedivý. 2015. Modeling of the Question Answering Task in the YodaQA System. In Josanne Mothe, Jacques Savoy, Jaap Kamps, Karen Pinel-Sauvagnat, Gareth Jones, Eric San Juan, Linda Capellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 222–228, Cham. Springer International Publishing.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *EMNLP*, pages 1533–1544. ACL.

E. Brill, S. Dumais, M. Banko, Eric Brill, Michele Banko, and Susan Dumais. 2002. An Analysis of the AskMSR Question-Answering System. In *Proceedings of EMNLP 2002*, January.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.

Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 845–855, Melbourne, Australia, July. Association for Computational Linguistics.

Danilo Croce, Alexandra Zelenanska, and Roberto Basili. 2018. Neural learning for question answering in italian. In Chiara Ghidini, Bernardo Magnini, Andrea Passerini, and Paolo Traverso, editors, *AI*IA 2018 – Advances in Artificial Intelligence*, pages 389–402, Cham. Springer International Publishing.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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.

David A. Ferrucci, Eric W. Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John M. Prager, Nico Schlaefer, and Christopher A. Welty. 2010. Building Watson: An Overview of the DeepQA Project. *AI Magazine*, 31(3):59–79.

Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2013. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524.

Sanda M. Harabagiu, Dan I. Moldovan, Marius Pasca, Rada Mihalcea, Mihai Surdeanu, Razvan C. Bunescu, Roxana Girju, Vasile Rus, and Paul Morarescu. 2000. FALCON: boosting knowledge for answer engines. In *Proceedings of The Ninth Text REtrieval Conference, TREC 2000, Gaithersburg, Maryland, USA, November 13-16, 2000*.

L. Hirschman and R. Gaizauskas. 2001. Natural language question answering: the view from here. *Natural Language Engineering*, 7(4):275–300.

Cody C. T. Kwok, Oren Etzioni, and Daniel S. Weld. 2001. Scaling question answering to the web. In *WWW*, pages 150–161.

Alexander H. Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In *EMNLP*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100.000+ Questions for Machine Comprehension of Text. *CoRR*, abs/1606.05250.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *International Conference on Acoustics, Speech and Signal Processing*, pages 5149–5152.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *5th International Conference on Learning Representations,*

ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net.

Huan Sun, Hao Ma, Wen tau Yih, Chen-Tse Tsai, Jingjing Liu, and Ming-Wei Chang. 2015. Open domain question answering via semantic enrichment. In *WWW*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.