Cross-lingual and cross-domain evaluation of Machine Reading Comprehension with Squad and CALOR-Quest corpora

Delphine Charlet, Géraldine Damnati, Frédéric Bechet, Gabriel Marzinotto, Johannes Heinecke

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5491–5497
Marseille, 11–16 May 2020

Cross-lingual and cross-domain evaluation of Machine Reading Comprehension with Squad and CALOR-Quest corpora

Delphine Charlet[1], Géraldine Damnati[1], Frédéric Béchet[2], Gabriel Marzinotto[1,2], Johannes Heinecke[1]
(1) Orange Labs, Lannion
(2) Aix-Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
(1) {first.last}@lis-lab.fr
(2) {first.last}@orange.com

Abstract

Machine Reading received recently a lot of attention thanks to both the availability of very large corpora such as SQuAD or MS MARCO containing triplets (document, question, answer), and the introduction of Transformer Language Models such as BERT which obtains excellent results, even matching human performance according to the SQuAD leaderboard. One of the key features of Transformer Models is their ability to be jointly trained across multiple languages, using a shared subword vocabulary, leading to the construction of cross-lingual lexical representations. This feature has been used recently to perform zero-shot cross-lingual experiments where a multilingual BERT model fine-tuned on a machine reading comprehension task exclusively for English was directly applied to Chinese and French documents with interesting performance. In this paper we study the cross-language and cross-domain capabilities of BERT on a Machine Reading Comprehension task on two corpora: SQuAD and a new French Machine Reading dataset, called **CALOR-QUEST** . The semantic annotation available on **CALOR-QUEST** allows us to give a detailed analysis on the kind of questions that are properly handled through the cross-language process.

Keywords: Machine Reading Comprehension, cross-lingual, FrameNet

1. Introduction

The Machine Reading Comprehension (MRC) task received recently much attention thanks to both the availability of very large corpora such as SQuAD or MS MARCO containing triplets (document, question, answer), and the introduction of Transformer Language Models such as BERT which obtains excellent results, even matching human performance according to the SQuAD leaderboard.

One of the key features of the Transformer Models is their ability to be jointly trained across multiple languages, using a shared subword vocabulary, leading to the construction of cross-lingual lexical representations. This feature has been used recently to perform zero-shot cross-lingual experiments where a multilingual BERT model fine-tuned on a MRC task exclusively for English was directly applied to Chinese and French documents with promising performance.

This study follows this path by comparing the impact of cross-language and cross-domain mismatch between training and testing corpora on a MRC task with BERT. Moreover, we propose to qualify each question according to a taxonomy and study which kind of questions are robust to cross-lingual transfer.

We perform our study on SQuAD as well as on a new French Machine Reading dataset, called **CALOR-QUEST** , which contains texts on other topics than SQuAD. One of the particularities of **CALOR-QUEST** is the fact that it was built on a semantically annotated corpus, with a Berkeley FrameNet model. For each question, a semantic description of the expected answer, as well as the question trigger is provided, allowing for a deeper result analysis and giving more insights about the behavior of BERT in a cross-lingual setting.

This paper is organised as follows: section 2 presents some related work on cross-lingual Machine Reading Comprehension evaluation with Transformers model; section 3.2 presents the **CALOR-QUEST** corpus with its semantic annotations; section 4 describes the taxonomy of questions derived from **CALOR-QUEST** and applied to SQuAD; section 5 reports the different cross-lingual and cross-domain experiments performed with detailed error analysis according to our question taxonomy.

2. Related work

Machine Reading Comprehension (MRC) is a task related to Question-Answering where questions are not generic in scope but are related to a particular document. Recently very large corpora (SQuAD (Rajpurkar et al., 2016), MS MARCO (Nguyen et al., 2016)) containing triplets (document, question, answer) were made available to the scientific community allowing to develop supervised methods based on deep neural networks with promising results. These methods need very large training corpora to be efficient, however such kind of data only exists for English at the moment.

Developing such resources for a new language requires a lot of effort, as presented in (He et al., 2018) for Chinese. Many methods have been proposed to help reducing this cost based on an automatic *translation* process between MRC resources in English and the target language (Asai et al., 2018; Lee and Lee, 2019).

In addition to methods performing a full translation of English corpora into a target language, methods have been proposed to directly perform online translation with a multilingual alignment process (Asai et al., 2018). One of the key issues with translation based methods, is the alignment between the source and target language when generating a text span answering a given question, as described in (Cui

et al., 2019). To overcome this problem several studies have proposed to take advantage of multilingual training of word representation in order to capture cross-lingual lexical representations. When training a machine reading model on a large source language corpus, and a much smaller target language corpus, this cross-lingual space allows the target model to benefit from the large source language training examples. Such approaches were proposed in (Cui et al., 2019; Kumar et al., 2019; Lee and Lee, 2019).

Another line of research consists in considering only multilingual lexical representations in a zero-shot cross-lingual setting where a machine reading model trained exclusively on a source language is applied to a target language for which no task-related data is available, an approach studied in (Siblini et al., 2019) and (Artetxe et al., 2019).

In this study we will investigate this zero-shot cross-lingual paradigm by applying a BERT multilingual model fine-tuned for an MRC task on a source language to an evaluation dataset on a target language. Moreover, we will compare cross-domain and cross-lingual performance and we will try to answer this question: which factor between language mismatch and domain mismatch has the strongest influence on the performances of an MRC task?

3. Machine reading Comprehension corpora description

3.1. SQUAD corpus

3.1.1. Dataset collection

The SQUAD corpus has been built from a collection of Wikipedia articles, selected from the top 10000 articles provided by the Project Nayuki's Wikipedia internal PageRanks. A subset of 536 articles has been randomly sampled from these articles. As a result, the collection has not been driven by any domain specific selection, but on the contrary the selection based on internal PageRank is likely to provide a variety of domains with general articles (the ones that are the more referred to in Wikipedia). We can thus argue the the SQUAD corpus reflects a "general knowledge" domain.

3.1.2. Question collection

In order to collect questions, the articles have been separated into articles and crowdworkers have been asked to ask and answer up to 5 questions for a given paragraph, within a time laps of 4 minutes. The answer has to be highlighted in the paragraph. As described into more details in (Rajpurkar et al., 2016), the crowdworkers were asked to use their own words but most questions were close to paraphrases with variations in syntactic constructions or small lexical variations with synonyms on the main verb for instance.

3.1.3. French-SQUAD subset

A translation into French of a subset of SQuAD proposed by (Asai et al., 2018). To do so, the authors trained an attention-based Neural Machine Translation model to translate questions and paragraphs. They proposed a method to align the start and end positions of the answer in the source language with a span of the context in the target language using attention weights. They have extracted with this approach several paragraphs and their corresponding question/answer pairs, resulting in 327 (question,answer, paragraph-question pairs over 48 articles. Thanks to the paragraph and question identifiers, we generated the same subset from the original English SQUAD corpus, allowing comparisons to be drawn from the same subset in two different languages.

3.2. CALOR-QUEST a machine reading corpus with semantic annotations

3.2.1. Dataset collection

One of the contributions of this work is to use a new French MRC corpus called **CALOR-QUEST** , developed from a corpus of encyclopedic documents annotated with semantic information (**CALOR-FRAME**) following the Berkeley Framenet paradigm described in (Marzinotto et al., 2018). The **CALOR-FRAME** corpus was initially built in order to alleviate Semantic Frame detection for the French language with two main purposes. The first one was to have a large amount of annotated examples for each Frame with all their possible frame Elements, with the deliberate choice to annotate only the most frequent Frames. As a result, the corpus contains 53 different Frames but around 26k occurrences of them along with around 57k Frame Element occurrences. The second purpose was to study the impact of domain change and style change. To this end the corpus was built by gathering encyclopedic articles from two thematic domains (`WW1` for First World War and `Arch` for Archeology and Ancient History) and 3 sources (`WP` Wikipedia, `V` for the Vikidia encyclopedia for children and `CT` for the Cliotext collection of historical documents) resulting in the 4 subcorpora described in Table 1.

collection	domain	source
V_antiq	Arch	V
WP_arch	Arch	WP
CT_WW1	WW1	CT
WP_WW1	WW1	WP

Table 1: CALOR corpus collection

As opposed to SQUAD articles collection approach that resulted in a "general knowledge" corpus, the **CALOR-QUEST** collection approach results in a specialized domain specific corpus.

3.2.2. Question collection

The approach followed to build **CALOR-QUEST** , described in (Béchet et al., 2019), is based on the use of semantic annotation in order to generate pairs of question/answer from an annotated document.

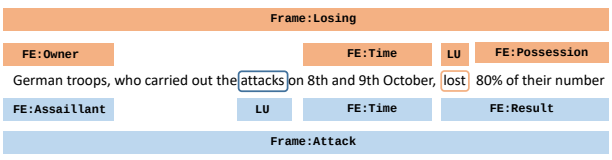

Figure 1: Example of sentence annotated with Frames in the **CALOR-FRAME** corpus

An example of annotated sentence[1] is given in figure 1. As one can see, a single sentence can be annotated with several frames, each of them triggered by a *Lexical Unit* (LU): `attacks` and `lost` for respectively the Frames `Attack` and `Losing`. To each Frame is associated a set of *Frame Element* (FE) which are the arguments of the semantic relation represented by the Frame. From such annotations, a *Question/Answer* corpus can be obtained. The method consists in producing, for every Frame occurrence f, a triplet (F, E, C) where F is the label of the frame f, E is one Frame Element of f and C (for Context) is the set of the other Frame Elements of f. Given a triplet (F, E, C), questions can be produced for which the answer is E.

Two processes have been used to generate the question/answer pairs of **CALOR-QUEST** : one based on an automatic generation process to obtain the training partition of **CALOR-QUEST** ; one based on a manual process for the validation/test partition.

The automatic generation process consists in using specific and generic patterns for generating questions from a triplet (F, E, C). Generic rules allow to produce a very large number of questions covering all possible questions a Frame could produce, without too much concern for the syntactic correctness of the questions produced. On the opposite, specific rules produce less questions but are closer to questions that one can naturally produce for a given Frame. In all cases the semantic validity of the questions generated is guaranteed, but not their syntactic validity.

The manual collection of *natural questions* is reduced to a validation/test set. Annotators had to write questions about a text paragraph directly from (F, E, C) triplets. The original sentence was not presented in order to leave more freedom for the annotator in her or his lexical and syntactic choices. Besides, the annotator can select any elements of the context to include in the question. The main advantage of this method is that it is possible to know, for each error made by an MRC system, which phenomenon was not well covered by the model. An example of the information provided to the annotator and the collected questions is given below:

Frame = `Hiding_objects`

- Context
  - **Agent** : `a Gallic militia leader`
  - **Hidden_object**: `a treasure`
  - **Hiding_place**: `in his Bassing farm`
- Answer
  - **Place**: `Moselle`

- Questions produced:
  - *In which region did the Gallic militia leader hide the treasure?*
  - *Where is the location of the Bassing farm in which the Gallic militia leader hid the treasure ?*

[1] the corpus is in French but the examples are translated in order to facilitate the reader's comprehension

With the proposed method, the resulting corpus CALOR-QUEST consists of about 300 documents in French, for which nearly 100 000 automatic question/answer pairs, and more than 1000 natural question/answer pairs are available. More detailed numbers per collection are given in table 2.

collection	#docs	#*natural* questions	#*generated* questions
V_antiq	61	274	4672
WP_arch	96	302	36259
CT_WW1	16	241	7502
WP_WW1	123	319	50971
total	**296**	**1136**	**99404**

Table 2: Description of CALOR-QUEST corpus

In the following, the **CALOR-QUEST** partition used to train models is made only of generated questions while the partition used for test is the *natural* question set.

4. Semantically-based taxonomy of questions

One of the key features of **CALOR-QUEST** is to provide a semantic representation of the answer for each question of the corpus. This representation of the expected answers is based on the Frame Element taxonomy used to annotate **CALOR-FRAME** . We propose to associate to each Frame Element a *question type* that can generalize the kind of answers represented by each Frame Element type. This taxonomy of questions is presented in section 4.1 for **CALOR-QUEST** and applied to SQuaD in section 4.2.

4.1. CALOR-QUEST question labelling

For each Frame Element, a *question type* can be assigned corresponding to the kind of question that could be asked to retrieve the Frame Element (FE). This question type is made of a *Wh-word* (interrogative determiners, adverbs, or pronouns) eventually followed by a specifier. We have performed this mapping association for the 53 Frames of the **CALOR-FRAME** corpus. For example, to the Frame `Hiding`, the FE `Hiding_place` is associated to the *Wh-word* `where`, `Hiding_object` is associated to `what` and `Agent` to `who_agent`. For "*who*" and "*what*", we make a distinction whether the FE corresponding to the answer is or is not the agent of the Frame. In the former case, the *question type* is `who_agent` or `what_agent`. In the latter case, it is simply `who` or `what`. We have chosen to make this distinction because some Frames can have several Frame Elements that can be retrieved by asking "what" or "who" question (eg. "*who attacks who?*"). Additionally, we assume that asking a question about the agent of a Frame can raise different issues than asking a question about its patient.

Eventually, all the 495 Frame Elements of the **CALOR-FRAME** corpus can be clustered into 64 question types. In this study they are used to categorize the questions of our test sets in order to study systems general performance along question types and to observe more deeply cross-domain and cross-language robustness of the models. The

CALOR-QUEST test partition covers all the Frames, but only a subset of 245 Frame Elements, corresponding to 38 different *question types* occur in this test partition. 20 of them occur more than 10 times, covering about 98% of the questions.

4.2. Bilingual subset of SQuAD question labelling

The French subset of SQUAD (Fr-sub-SQUAD) will be used in our study both in its French version (`Fr-sub-SQuAD`) and in its corresponding original English version (`En-sub-SQuAD`) to support our experimental observations. We have manually labelled this subset of questions with the same *question type* taxonomy as the one obtained from **CALOR-QUEST** . We want to point out that this manual annotation process is not based on the *Wh-word* used in the questions, but rather on the semantic role of their answers. For instance, all questions for which the answer is a place are labelled `where` whatever the formulation of the question (e.g. *"What venue did Super Bowl 50 take place in?"*).

4.3. Question type distribution

Figure 2 presents the distribution of the 20 most frequent question types in both CALOR and sub-SQuAD corpora. In both cases the majority class is the category `what` but the proportion is higher for sub-SQuAD. Questions related to amounts (`how_much`) are much more represented in the sub-SQuAD corpus while the category `who_agent` is more represented in **CALOR-QUEST** . The other categories are rather similarly distributed. Adverbial phrases on the right of the histogram correspond to a minority of questions in **CALOR-QUEST** but present some interesting properties in terms of models robustness.

5. Experiments

5.1. Experimental set up

We use for our MRC system a fine-tuned version of BERT multilingual model: *multi_cased_L-12_H-768_A-12* (Devlin et al., 2018)[2], with default hyperparameters. In order to reproduce the same conditions as the SQuAD corpus, we cut the CALOR documents into paragraphs whose lengths are close to the average paragraph length of SQuAD (around 120 tokens): starting at the beginning of each document, we look for the next end of sentence marker after 120 tokens. This constitutes the first paragraph on which the MRC system will be applied. Then the process is repeated on the text starting at the next sentence in the document.
The evaluation is done with SQuAD's evaluation script[3], customized for French (removing french articles in the normalization process, instead of English articles) when evaluating French corpora. In this evaluation set-up, "*F1*" is the average F-measure per question, where for each question a precision/recall performance is measured between the predicted and ground-truth sets of tokens in answer spans.

[2] `https://github.com/google-research/bert/blob/master/run_squad.py`)

[3] `https://github.com/allenai/bi-att-flow/blob/master/squad/evaluate-v1.1.py`

Models that are claimed to be trained on SQuAD in this paper are trained on the standard training set of SQuADv1.1. Training on the CALOR corpus is done on a selected sample set of 30K generated questions. The sampling strategy is motivated by the objective of covering with at least one question all the answers of the corpus. When an answer has several candidate questions, the priority is given to questions generated with specific rules (which are closer to natural questions). Finally the set is completed with questions produced by generic rules in order to reach a 30k training size. Experiments, not reported here, have shown that adding many variants of generic questions for each answer does not improve performances, and may even decrease them.

5.2. Comparing cross-domain and cross-language robustness

Table 3 compares performances in terms of F1 of a BERT model trained on SQuAD or trained on CALOR for various test sets. For the first two lines, the test sets differ both from language and domain (even though both SQuAD and CALOR are intrinsically multi-domain, we use the term "domain" to refer the switch in corpus collection condition).

Training Corpus Train. Size (Lang.)	SQuAD 87599 (En)	CALOR 30000 (Fr)
Fr-CALOR-Quest	62.1	82.0
En-SQuAD	87.9	58.3
En-sub-SQuAD	88.8	58.7
Fr-sub-SQuAD	79.6	61.8

Table 3: F1 on various test sets in French and in English with models trained on SQuAD (trainv1.1) or **CALOR-QUEST**

We can observe the impact of train/test mismatch. However, simply observing those two lines is not enough to conclude if language or domain switch are responsible for the drop in performances. The third and fourth line provide some interesting insights. If the model trained on French data yield slightly better performances on Fr-sub-SQuAD than on En-sub-SQuAD (61.8 *vs* 58.7), the model trained on English SQuAD outperforms the model trained on French CALOR for both languages. Even if training corpora don't have the same size, it is particularly striking to observe that a model trained in similar conditions but on a different language can yield 79.6 F1 while a model trained on French data in different conditions provides a 17.8 pts absolute drop in average F1. These results tend to show that Machine Reading Comprehension models based on multilingual BERT contextual representations are more robust to language variations than to domain variations.
Figure 3 shows detailed performances of models trained on CALOR or SQuAD on the **CALOR-QUEST** *natural* questions test set, depending on the question type labelled as presented in section 4.1. Categories are ranked according to decreasing average F1 score for the model trained on SQuAD. The count of each question type is given in the x axis labels.

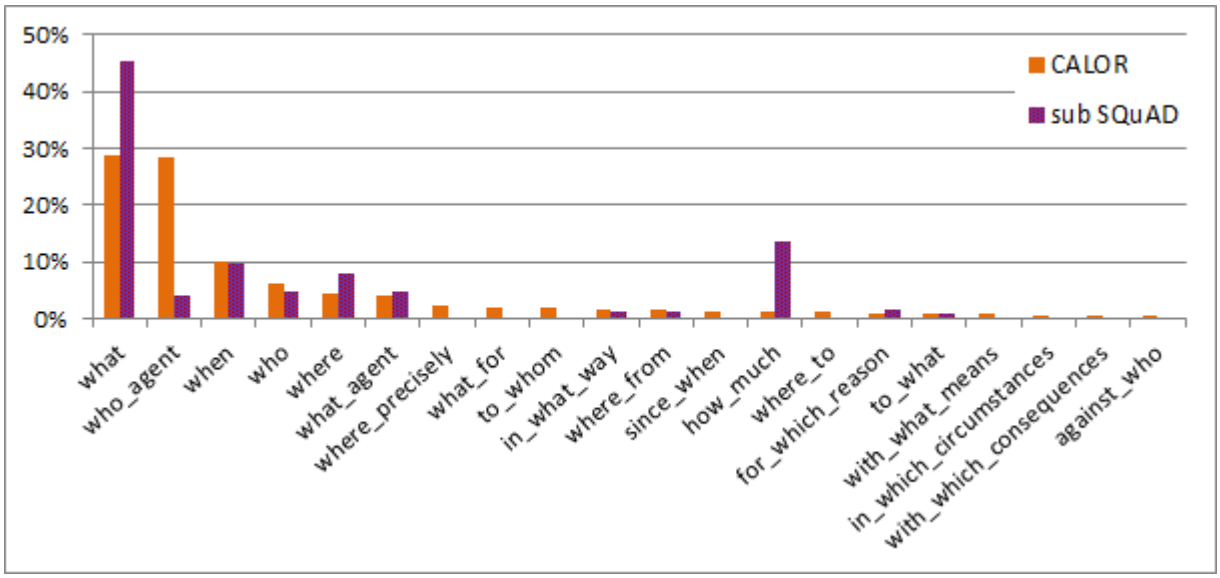

Figure 2: distribution of question types for CALOR and the bilingual subset of SQuAD

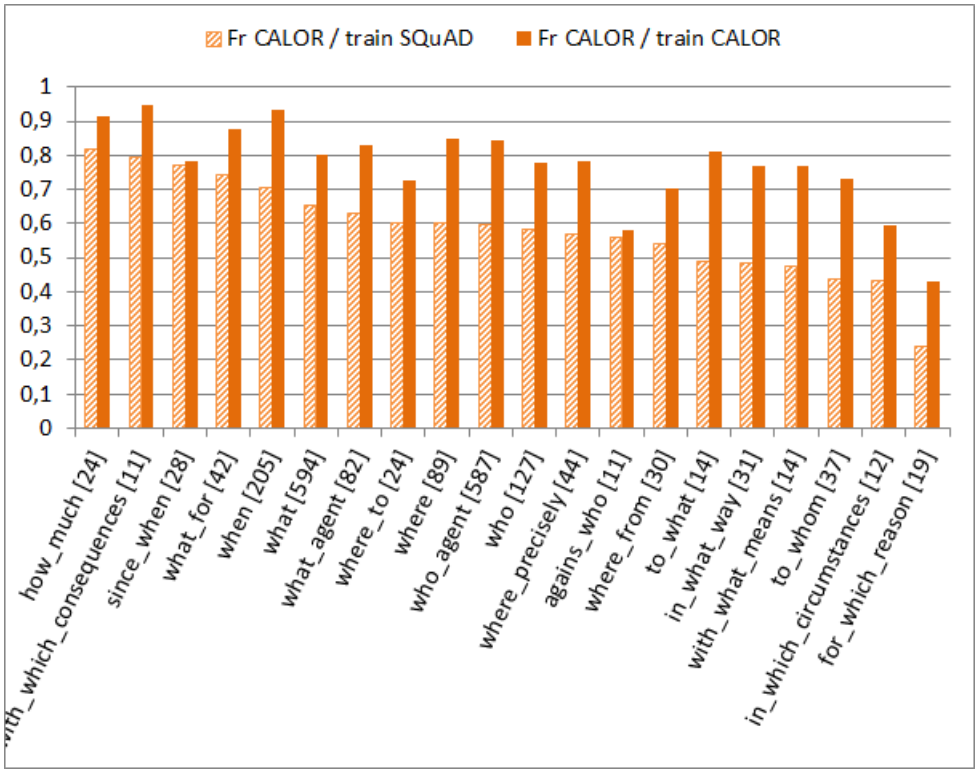

Figure 3: Detailed performance by question type on the **CALOR-QUEST** test set

The first observation is that performance is not uniform among question types, with performance ranging from 24% to 82% for SQuAD model, and from 42% to 95% for CALOR model. All question types improve their performance when the model is trained on the dedicated CALOR corpus, with a relative improvement ranging from 2% (`since_when` to 78% (`for_which_reason`). Even if the conclusions must be tempered by the small amount of examples for these labels, the observation could be explained by the fact that `since_when` answers are strongly correlated to a stable syntactic construction with a particular preposition (since or *depuis* in French) followed by a date. On the other hand, (`for_which_reason`) answers are more likely to occur with a large variety of formulations, making it more difficult for models to generalize. The question types which perform almost the same, whatever the model, are `against_who` and `since_when` but, here again, the small size of these sets does not allow to draw definitive conclusions. The question types with the largest improvement (around 60% and more) are: `for_which_reason`, `to_whom`, `to_what`, `with_what_means` and `in_what_way`. `How_much`

question type performs well (¿80%) already with SQuAD model, and its improvement with dedicated model is moderated. Finally, the worst performing question-type is the same for both models: `for_which_reason`, even though there is a noticeable improvement with CALOR corpus.
The next figure focuses on the SQuAD bilingual test set. It is important to note from table 3 that the overall results on full En-SQuAD and on En-sub-SQuAD are comparable and for the rest of this section we will study in depth models performance on sub-SQuAD only. Hence figure 4 shows performance for the question-types with more than 10 occurrences (covering more than 90% of the sub-SQuAD questions) for different versions of the test set (French or English) and for different models: the model trained on SQuAD and the model trained on CALOR.
It can be observed that most of the time, results obtained on different question types are comparable across languages for a given model, except for `who_agent` and `what_agent` where these results are less stable. However, the difference due to model training conditions is very important. Hence, training on the CALOR French corpus, which is not suited for the thematic domains of SQuAD questions, does not help answering French questions and performance remains at the level of the English ones. That pleads for the intrinsic multilingual capacity of BERT, but domain specific data still enables to get the best performance. In other words the language effect has less impact than the domain shift effect.

5.3. Mixing training corpora

In this section we want to observe the impact of mixing training data from different languages and different domains. Hence table 4 contains a third column where the model has been trained on a hybrid training corpus composed of the **CALOR-QUEST** 30k French question/paragraph pairs and a randomly sampled subset of 60k English SQuAD question/paragraph pairs. Sub-sampling the SQuAD corpus was achieved in order to obtain equivalent corpora in terms of training size (around 90k question/paragraph pairs).

Training Corpus Train. Size (Lang.)	SQuAD 87599 (En)	CALOR 30000 (Fr)	mixed 30k(Fr)) +60k(En)
Fr-CALOR-Quest	62.1	82.0	82.6
En-SQuAD	87.9	58.3	87.6
En-sub-SQuAD	88.8	58.7	87.7
Fr-sub-SQuAD	79.6	61.8	75.4

Table 4: F1 on various test sets with models trained on SQuAD (trainv1.1), CALOR-Quest and a mix of both

The first row shows that performance on the Fr-CALOR-Quest test set is only slightly impacted by training data augmentation (82.6 F1 with mixed training corpus against 82.0 with `Calor` only. A more detailed analysis showed however that some question types are not impacted by corpus augmentation (`what`, `when`, `how_much`, `to_what` and `what_for`) while some are positively impacted (+10% relative for `to_whom` and +36% relative for `against_who`) and some on the contrary are negatively impacted (-15% relative for `in_what_way` and -30% relative for `for_which_reason`). The second and third rows yields similar conclusions: training on a multilingual corpus has little impact on overall performance for English test sets. The last row however confirms the preceding conclusions regarding the larger influence of domain against language. For the French subset of SQuAD adding training data from the same domain in English is more efficient (79.6 average F1) than adding training data in French from another domain.

6. Conclusion

This paper presented a study on the cross-language and cross-domain capabilities of BERT on a Machine Reading Comprehension task. We have described a new French Machine Reading dataset, called **CALOR-QUEST** . The semantic annotation available on **CALOR-QUEST** allows us to give a detailed analysis on the types of questions that are properly handled through the cross-language process. Thanks to a semantically motivated taxonomy of (question, paragraph, answer) triplets, we were able to highlight which question types are intrinsically more error prone and which are more sensitive to language or domain shift. Finally, the use of a subset of the SQuAD test corpus available in both French and English allowed further observations to be drown. One of the main conclusion of this work is that with multilingual contextual word representations, the language effect has less impact than the domain shift effect.

7. Bibliographical References

Artetxe, M., Ruder, S., and Yogatama, D. (2019). On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*.

Asai, A., Eriguchi, A., Hashimoto, K., and Tsuruoka, Y. (2018). Multilingual extractive reading comprehension by runtime machine translation. *CoRR*, abs/1809.03275.

Béchet, F., Aloui, C., Charlet, D., Damnati, G., Heinecke, J., Nasr, A., and Herledan, F. (2019). CALOR-QUEST : generating a training corpus for Machine Reading Comprehension models from shallow semantic annotations. In *MRQA: Machine Reading for Question Answering - Workshop at EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing*, Hong Kong, China, November.

Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., and Hu, G. (2019). Cross-lingual machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1586–1595.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

He, W., Liu, K., Liu, J., Lyu, Y., Zhao, S., Xiao, X., Liu, Y., Wang, Y., Wu, H., She, Q., et al. (2018). Dureader: a chinese machine reading comprehension dataset from real-world applications. *ACL 2018*, page 37.

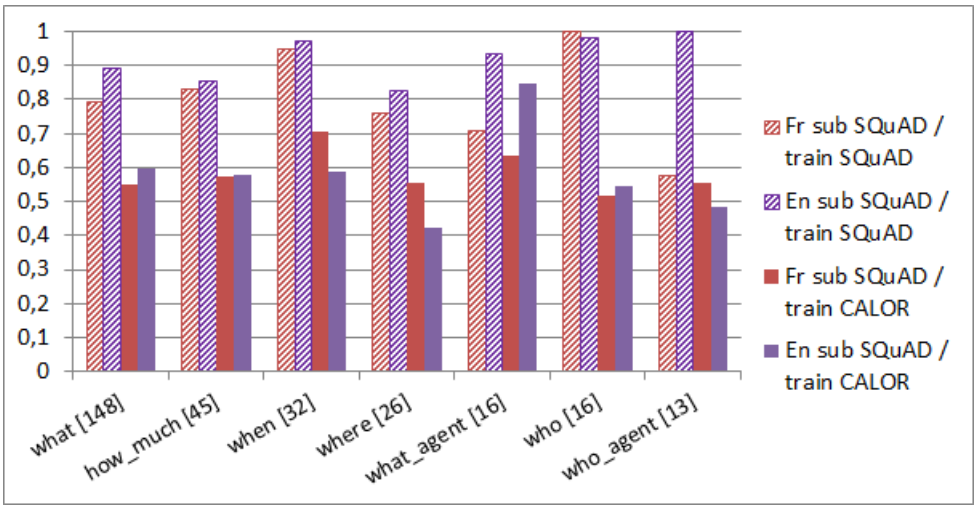

Figure 4: Result on bilingual subset of SQUAD corpus

Kumar, V., Joshi, N., Mukherjee, A., Ramakrishnan, G., and Jyothi, P. (2019). Cross-lingual training for automatic question generation. *CoRR*, abs/1906.02525.

Lee, C.-H. and Lee, H.-Y. (2019). Cross-lingual transfer learning for question answering. *arXiv preprint arXiv:1907.06042*.

Marzinotto, G., Auguste, J., Bechet, F., Damnati, G., and Nasr, A. (2018). Semantic frame parsing for information extraction : the calor corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Language Resource Association.

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., and Deng, L. (2016). Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.

Siblini, W., Pasqual, C., Lavielle, A., and Cauchois, C. (2019). Multilingual question answering from formatted text applied to conversational agents. *arXiv preprint arXiv:1910.04659*.