etalab gouv.fr

reciTAL.

Project PIAF

PROTOCOL

1 Motivation

NLP applications are naturally bound to language specificities, a fact that highlights the asymmetry in the availability of NLP datasets, with the great majority targeting the English language. The significance of this issue has been recently acknowledged by a resolution of the EU Parliament[1].

Over the years, the community has produced several resources to tackle tasks we call, for simplicity, *upstream* (such as Part-Of-Speech tagging, Dependency Parsing, etc.), targeting multiple languages and enabling the construction of effective automated systems. Still, for those tasks we refer to as *downstream*, i.e. those which enable the development of value-added end products such as Question Answering or Conversational Agents, the current state-of-the-art approaches based on Deep Learning methodologies require massive amounts of annotated data which are almost exclusively available in English. Notable exceptions are tasks for which such data is available, such as Machine Translation, for which abundant parallel corpora have been built from resources such as the European parliamentary proceedings, or Language Modeling which can be tackled in a semi-supervised manner hence only requiring massive amounts of text in the target language(s).

A significant performance gap is observed when trying to build a French QA system from English data via automatic translation to French. It can be hypothesized that such gap is at least partially explained by shortcomings of the Machine Translation models used. Nonetheless, an open research question is the extent to which the performance of state-of-the-art models can transfer to other languages than English. Recent research [1] on Chinese language (arguably, the other high-resource language[2] along with English) indicates that, while translation-based method and multilingual approaches can obtain reasonable performances, there exist a large margin for improvements. To allow the research community to tackle those issues, it is thus desirable to obtain comparable data for the task, natively for the language of choice. Efforts in this direction have focused on Chinese [2], Korean [3, 4], and – to a limited extent – German [5].

Aiming to fill this gap for French language, we present in this document the protocol we followed to collect a French QA dataset similar to the popular "Stanford Question Answering Dataset" (SQuAD) [6].

2 Collection Setting

The data collection will not rely on *gig-economy* platforms (e.g. Mechanical Turk), and will thus take place during several events. Besides the value for both research and industry of the data that will be collected, this project aims at engaging the French-speaking community as well as contributing to the awareness of public administrations with regards to Artificial Intelligence (AI). At the same time, it will shed light on how the latest generation models are trained, and the biases they encode. As a side effect, it will provide information on the attractiveness and effectiveness of a public collaborative effort. For this data collection effort, we rely on the French version of Wikipedia: we select relevant articles, segment those in smaller paragraphs, and collect sets of question/answer pairs corresponding to those paragraphs.

Two main collection settings are envisioned: the first, *certified*, wherein participation to the collection is restricted to employees of the entities involved; the second, *crowd-sourcing*, will be open for online participation.

The dataset produced will include three splits, namely:

- *training*, accounting for 80% of the source articles;
- *validation*, accounting for 10% of the source articles;
- *test*, accounting for 10% of the source articles.

[1] European Parliament resolution of 11 September 2018 on language equality in the digital age. `http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P8-TA-2018-0332+0+DOC+XML+V0//EN`

[2] In terms of publicly available text and corpora.

We refer to the union of *validation* and *test* splits as *development* data.

As the *certified* setting is by definition more controlled and the quality of the data is expected to be higher in this setting, we will rely on it for both *development* and *training* data collection. The *crowd-sourcing* setting will only be leveraged for the latter. The rationale behind this choice is that evaluation data (i.e. *development* and *test* sets) should be of high quality, while the presence of lower quality samples in the much larger *training* set does not hinder building successful models (arguably, it rather helps obtaining more robust models).

3 Continuous Evaluation

It is of utmost importance to drive the users to produce challenging questions. In the SQuAD collection interface, the user was reminded to avoid using the same words/phrases as the paragraph while writing a question, via a text message displayed on screen. In our collection interface, we will give examples of good and bad questions to the user. The evaluation methods to measure the question quality are detailed in this section.

In order to check the quality of the collected data, both manual and automatic evaluations will be carried out throughout the data collection process. Those will be performed on a rolling basis, as the collected *development* set grows. The manual and automatic evaluations are explained and illutrated by exemples in this section.

3.1 Manual

We subsample the data and evaluate triplets ($\{Paragraph, Question, Answer\}$) according to the reasoning required:

- *lexical variation*: the lower the number of common words between the paragraph and the question, the highest the measure of lexical variation. For example, consider the sentence : *"During the Gaulish period, Nantes belonged to the Namnetes people."*; the question *"To whom did Nantes belong during the Gaulish period ?"* has a lower lexical variation than *"Who inhabitated Nantes before the Romans came ?"*.

- *syntactic variation*: this measure can be computed automatically and is explained in the next section.

- *multiple sentence reasoning*: it is necessary to combine information from multiple sentences to answer the question. The question *"What is today's name of the Roman city Condevincum ?"* on the following pargaraph *"During the Gaulish period, Nantes belonged to the Namnetes people. [...] The Romans latinized its name to Condevincum."* needs the two sentences to be answered with *"Nantes"*;

- *ambiguity* : the question is unclear and/or has multiple possible answers in the text. For example, in the following paragraph : *"Around 490, the Franks under Clovis I captured Nantes (alongside eastern Brittany) from the Visigoths after a sixty-day siege;it was used as a stronghold against the Bretons."*, the question "*Who lived in Nantes during the fifth century*" has no unique answer.

For this, will target a very small subsample of the *development* set (in the original SQuAD paper, this amounts to a total of 192 questions, i.e. 4 per article).

3.2 Automatic

We will use scripts to automatically evaluate syntactic divergence between questions and corresponding answers. This will be separately applied to each dataset split. Consistently with the original SQuAD study, we compute a measure for syntactic divergence as the edit distance between the unlexicalized dependency paths in the question and the sentence containing the answer. This provides a proxy to understand the complexity of the questions posed, i.e. the amount of reasoning required to answer them. For instance, assume a question *"Who went to Wittenberg to hear Luther speak?"* and a corresponding text *"Students thronged to Wittenberg to hear Luther speak."*: the comparison of the dependency paths of the two phrases shows that they are equal; in other words the structure of the two phrases is the same and the amount of reasoning required to answer *"Who"* with *"Luther"* is minimal.

Conversely, consider the question *"What impact did the high school education movement have on the presence of skilled workers?"* with a corresponding text *"During the mass high school education movement from 1910 1940 , there was an increase in skilled workers"*: in this case, the higher score for syntactic divergence indicates that more complex reasoning is required to provide an answer.

Further, we will incrementally measure performance of state-of-the-art multi-lingual QA systems, under different experimental setups. The results of this latter evaluation stage will eventually trigger moving to a subsequent stage of the data collection procedure.

4 Data Collection

The data collection will start targeting a dataset comparable to SQuADv1.1, for which the great majority of questions are expected to have an exact answer in the corresponding paragraph[3]. The collection effort will focus on obtaining a reliable evaluation dataset first, at a scale of about 20k samples. We subsequently target to collect about 80k training samples.

We envision several stages of the data collection effort, which we sequentially detail below. Depending on the continuous evaluation described in 3.2 we will eventually trigger the switch to subsequent stages of the data collection, designed to increase the difficulty of the dataset.

4.1 Base Collection (SQuADv1.1)

This mode serves as a base for all envisioned collection modes, and aims at collecting data comparable to SQuADv1.1. To this end, two versions of the UI are required, and described in sections 4.1.3 and 4.1.4. We assume each participant is associated to a unique identifier that can link them to a *certified* or *crowdsourcing* status.

4.1.1 Source Data Selection

Consistently with SQuAD, the Wikipedia articles to use as source for the question-answering task will be selected according to the following methodology, with the required motivation as explained in 7.1. First, the PageRank[4] centrality score will be computed for each article in the French Wikipedia, in order to rank the articles and subsequently select the most relevant ones[5]. From each, finally, we will select the individual paragraphs, stripping away images, figures, tables, and discarding paragraphs shorter than 500 characters. The articles will be randomly partitioned, as mentioned above, in a 80-10-10 split, respectively for train-dev-test.

4.1.2 Guidelines Screen

After logging in, the user is shown a sample paragraph, as well as examples of good and bad questions/answers on that paragraph. An explanation for why they were categorized is also displayed.

4.1.3 Training data collection

The users are prompted with a paragraph selected among those which have no associated question/answer pairs yet. In Fig. 1 we report a screenshot of the original UI used for collecting the data in this mode. Copy paste functionalities should be disabled: a user should not be able to copy text from the paragraph box into the question boxes. When the "select answer" button is clicked, the user can select/highlight a span in the paragraph – which will be recorded as the answer to the corresponding generated question. The annotator can either choose to submit

[3] for SQuADv1.1, the amount of questions marked as unanswerable in the *development* set amounts to only 2.6%

[4] `https://en.wikipedia.org/wiki/PageRank`

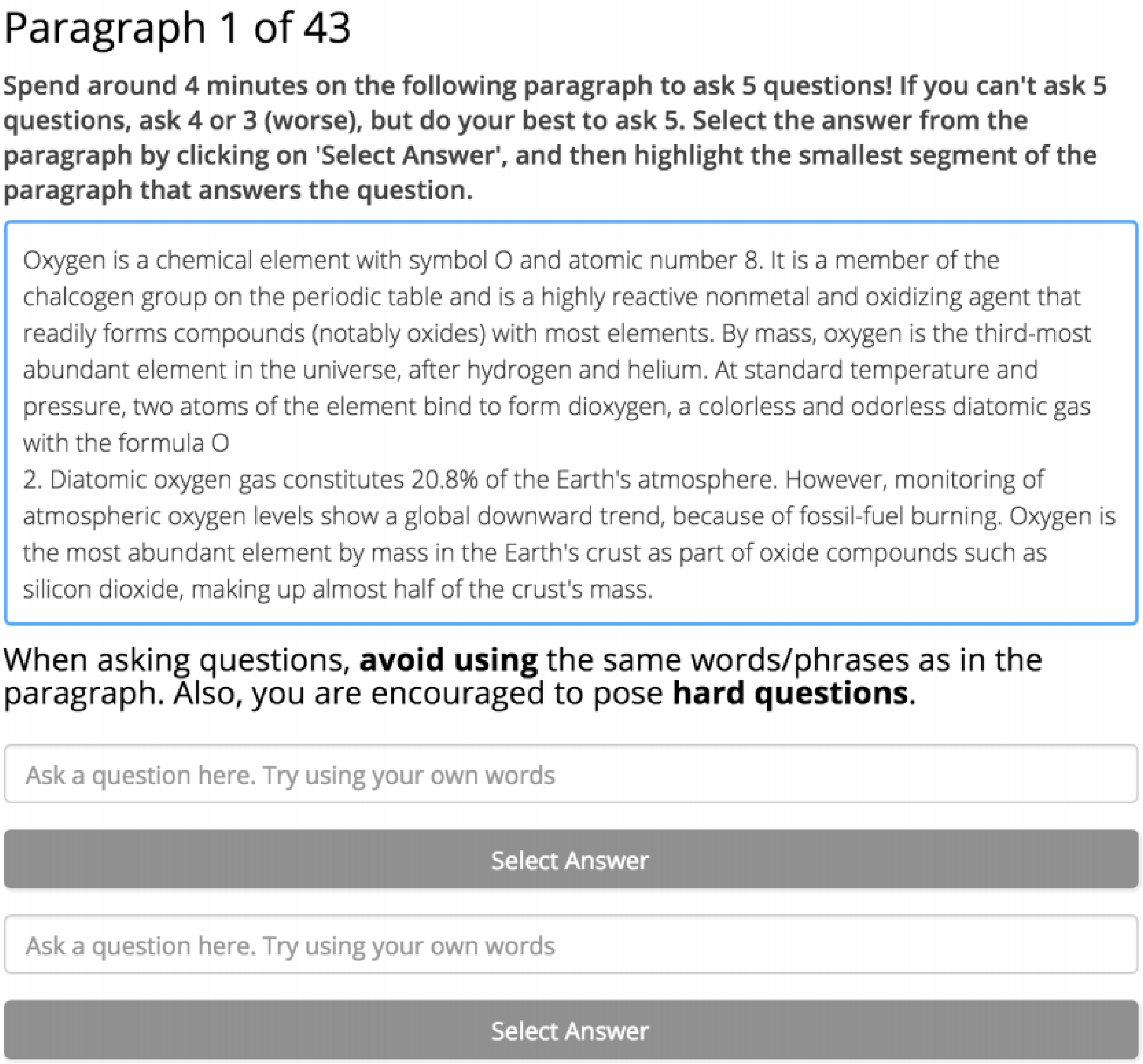
[5] We will use, with the required modifications, the code made available by Project Nayuki, `https://www.nayuki.io/page/computing-wikipedias-internal-PageRanks`

the current paragraph along with the generated question/answer pairs, or to continue generating questions. The UI should drive the annotator to generate five question/answer pairs per paragraph before submitting.

4.1.4 Evaluation data collection

In order to make evaluation data more robust, as well as to obtain measures on human performance on the task, additional answers are collected for each sample in *validation* and *test* splits (i.e., the development set).

To this end, a user is first prompted with the instructions/guidelines screen for this task. Then, the user is only shown a paragraph and a question previously generated by another user. His task is then to select the answer to the generated question, or mark the question as *unanswerable* otherwise (in SQuADv1.1 this was assumed to be the case when the user selected no answer at this stage).



Paragraph 1 of 43

Spend around 4 minutes on the following paragraph to ask 5 questions! If you can't ask 5 questions, ask 4 or 3 (worse), but do your best to ask 5. Select the answer from the paragraph by clicking on 'Select Answer', and then highlight the smallest segment of the paragraph that answers the question.

Oxygen is a chemical element with symbol O and atomic number 8. It is a member of the chalcogen group on the periodic table and is a highly reactive nonmetal and oxidizing agent that readily forms compounds (notably oxides) with most elements. By mass, oxygen is the third-most abundant element in the universe, after hydrogen and helium. At standard temperature and pressure, two atoms of the element bind to form dioxygen, a colorless and odorless diatomic gas with the formula O
2. Diatomic oxygen gas constitutes 20.8% of the Earth's atmosphere. However, monitoring of atmospheric oxygen levels show a global downward trend, because of fossil-fuel burning. Oxygen is the most abundant element by mass in the Earth's crust as part of oxide compounds such as silicon dioxide, making up almost half of the crust's mass.

When asking questions, **avoid using** the same words/phrases as in the paragraph. Also, you are encouraged to pose **hard questions**.

Ask a question here. Try using your own words
Select Answer
Ask a question here. Try using your own words
Select Answer

Figure 1: UI for collecting samples for SQuADv1.1

4.2 Data Format

All datasets will share a commmon top-level structure - consistent with the SQuAD format:

```
{'data': [] # type: list -- list of wikipedia articles
 'version': '' # type: string, version of the dataset (e.g. '1.1')
}
```

As seen above, the `data` key maps to a list of wikipedia articles, structures as:

```
{'title': '' # type: string -- title of the wikipedia page
 'paragraphs': [] # type: list -- list of paragraphs in the wikipedia page
}
```

The structure of the elements in the `paragraphs` list is consistent among training, development, and test splits. The output format for training data paragraphs follows:

```
{'context': 'Architecturally, the school has a Catholic character.
            Atop the Main Building\'s gold dome is a golden statue of the
            Virgin Mary. Immediately in front of the Main Building and facing
            it, is a copper statue of Christ with arms upraised with the legend
            "Venite Ad Me Omnes". Next to the Main Building is the Basilica of
            the Sacred Heart. Immediately behind the basilica is the Grotto, a
            Marian place of prayer and reflection. It is a replica of the
            grotto at Lourdes, France where the Virgin Mary reputedly appeared
            to Saint Bernadette Soubirous in 1858. At the end of the main drive
            (and in a direct line that connects through 3 statues and the Gold
            Dome), is a simple, modern stone statue of Mary.',
 'qas': [
        {'answers': [{'answer_start': 515, 'text': 'Saint Bernadette  Soubirous'}],
         'question': 'To whom did the Virgin Mary allegedly appear in 1858 in  Lourdes France?',
         'id': '5733be284776f41900661182'
        },
        {'answers': [{'answer_start': 188, 'text': 'a copper statue of Christ'}],
         'question': 'What is in front of the Notre Dame Main Building?',
         'id': '5733be284776f4190066117f'
        },
        {'answers': [{'answer_start': 279, 'text': 'the Main Building'}],
         'question': 'The Basilica of the Sacred heart at Notre Dame is beside to which structure?',
         'id': '5733be284776f41900661180'
        },
        {'answers': [{'answer_start': 381, 'text': 'a Marian place of prayer and reflection'}],
         'question': 'What is the Grotto at Notre Dame?',
         'id': '5733be284776f41900661181'
        },
        {'answers': [{'answer_start': 92, 'text': 'a golden statue of the Virgin Mary'}],
         'question': 'What sits on top of the Main Building at Notre Dame?',
         'id': '5733be284776f4190066117e'
        }]}
```

For each question, training samples can only have one answer, while evaluation samples may have several answers.

setting	*samples per minute*	*man/h for 10k samples*	*samples per session*
base (training)	1.25	133.3	6000
base (additional/dev)	2.5	66.6 (x n_req)	$12000/n\_req$
unanswerable	1.4	233.3	6720

Table 1: Estimated time required for each setting. n_req is the minimal amount of additional answers required (in SQuADv1.1, $n\_req = 2$)

5 Protocol Switches

Following automatic evaluation, we should be able to switch at any moment to a slightly different collection protocol, in order to obtain increasingly challenging questions. We will trigger such switch in case the automatic evaluation on the development set suggests that results are comparable with those on SQuAD. Samples generated within a specific switch should be marked as such via a specific flag/marker.

5.1 Unanswerable Questions

Under this setting, we will collect *plausible* yet *unanswerable* samples (the answer is not found in the context paragraph), targeting in terms of size 50% of the already collected data at most. This is consistent with SQuAD version 2.0; the UI used for SQuADv2.0 is shown in Figure 2. Besides the changes in guidelines, the only difference, in terms of UI, with respect to that shown in 1 lies in showing “Questions for inspiration": those are the *answerable* questions already collected in the previous collection stage.

In terms of format, this setting requires adding two new fields (*plausible_answers* and *is_impossible* to elements in the *qas* list.

```
{'plausible_answers': [{'text': 'Normandy', 'answer_start': 137}],
 'question': 'What is France a region of?',
 'id': '5ad39d53604f3c001a3fe8d2',
 'answers': [],
 'is_impossible': True}
```

5.2 Adversarial Baseline

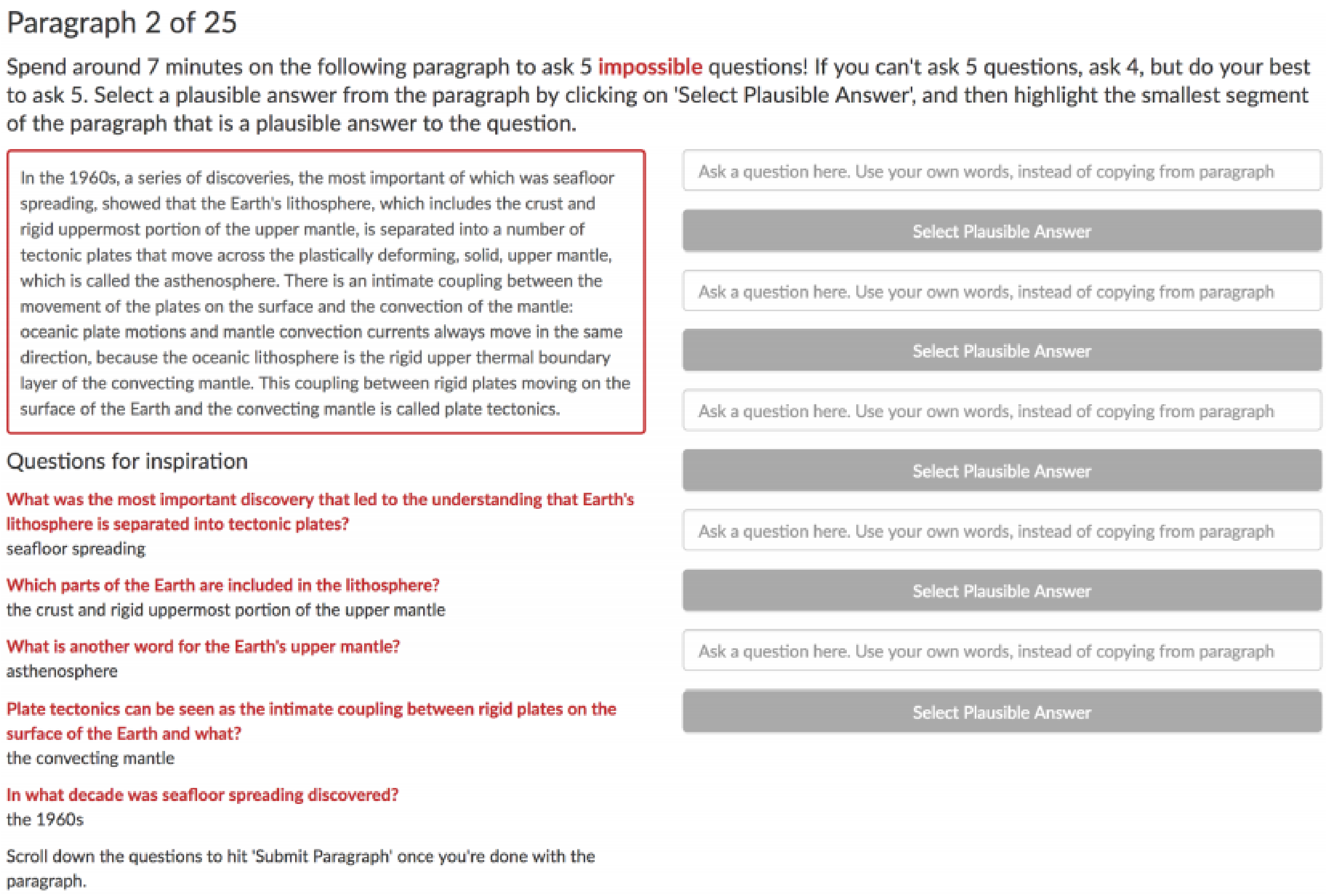
In order to incentivize crowd-workers to produce challenging questions, we will provide a pre-trained QA service endpoint. The users will be asked to submit a question only if the system is unable to answer it correctly. As soon as a user types a question, the QA service will be invoked; the answer it returns will be highlighted in the paragraph and the user will be asked to submit it only if such answer is incorrect.

As a baseline, we will use the pre-trained multilingual BERT model, fine-tuned on the English SQuAD training data. This model will be evaluated both on the English and French dev sets to measure and compare the performances. It will also allow to efficiently measure the impact of data-augmentation techniques (e.g. automatic translation of the training samples).

6 Timing Estimations

In Table 1 we report the timing estimates to gather the corpus under the SQuAD-based scenarios (data from SQuAD reports is used).

Assuming 20 volunteers per session, with a session consisting of 4-hours of effective annotation time, we can expect to collect an amount of evaluation data comparable to SQuAD (20k) in 4 or 5 sessions (for $n\_req = 2$), around 320-400 hours.



Paragraph 2 of 25

Spend around 7 minutes on the following paragraph to ask 5 **impossible** questions! If you can't ask 5 questions, ask 4, but do your best to ask 5. Select a plausible answer from the paragraph by clicking on 'Select Plausible Answer', and then highlight the smallest segment of the paragraph that is a plausible answer to the question.

In the 1960s, a series of discoveries, the most important of which was seafloor spreading, showed that the Earth's lithosphere, which includes the crust and rigid uppermost portion of the upper mantle, is separated into a number of tectonic plates that move across the plastically deforming, solid, upper mantle, which is called the asthenosphere. There is an intimate coupling between the movement of the plates on the surface and the convection of the mantle: oceanic plate motions and mantle convection currents always move in the same direction, because the oceanic lithosphere is the rigid upper thermal boundary layer of the convecting mantle. This coupling between rigid plates moving on the surface of the Earth and the convecting mantle is called plate tectonics.

Questions for inspiration

What was the most important discovery that led to the understanding that Earth's lithosphere is separated into tectonic plates?
seafloor spreading

Which parts of the Earth are included in the lithosphere?
the crust and rigid uppermost portion of the upper mantle

What is another word for the Earth's upper mantle?
asthenosphere

Plate tectonics can be seen as the intimate coupling between rigid plates on the surface of the Earth and what?
the convecting mantle

In what decade was seafloor spreading discovered?
the 1960s

Scroll down the questions to hit 'Submit Paragraph' once you're done with the paragraph.

Ask a question here. Use your own words, instead of copying from paragraph
Select Plausible Answer
Ask a question here. Use your own words, instead of copying from paragraph
Select Plausible Answer
Ask a question here. Use your own words, instead of copying from paragraph
Select Plausible Answer
Ask a question here. Use your own words, instead of copying from paragraph
Select Plausible Answer
Ask a question here. Use your own words, instead of copying from paragraph
Select Plausible Answer

Figure 2: UI for collecting *unanswerable* questions for SQuADv2.0

7 Workflow

7.1 Offline Data Selection

In SQuAD, the authors used the highest-ranking, in terms of PageRank, 10k articles from the English Wikipedia. Nonetheless, when applying the same threshold on the French version, we noticed significant differences in terms of structural properties between French and English Wikipedia. For instance, we observed the massive presence of pages referring to *years* on the French version, a characteristic that the top-10k subsample of the English Wikipedia does not seem to have. After manual inspection, the editing practices seem to differ between the French and English communities: while the latter do not link all *year* mentions to the actual page dedicated to that year, the French Wikipedia editors seem to systematically do so — a fact that boosts the PageRank score of such articles and therefore explains their abundant presence in the French subsample.

Furthermore, given the significant presence of Wikipedia articles not exploitable for our goals (e.g. drafts, disambiguation pages), we operate as follows:

- gather the top-25k articles in terms of PageRank;
- discard the unexploitable articles;
- set a min-max char limit on the paragraph length ($min = 500; max = 1000$);
- filter out articles with less than 5 paragraphs;
- apply the train/dev/test splitting procedure.

Comparing to the English SQuAD, we thus expect to obtain annotated QA data on more articles, with less paragraphs per article, and a comparable length. This compromise seems reasonable: the main drawback is the availability of fewer paragraphs per article, but since this factor is mostly relevant for negative sampling[6], a $5 : 1$ ratio seems acceptable.

For each article, a unique id is computed – e.g. by taking the hash of the title. Each paragraph is assigned a unique id composed by the id of the article it belongs to, plus a sequential number corresponding to its position within the article.

7.2 Online Paragraph Selection

The algorithm for selecting which paragraph should be displayed to a volunteer is defined as follows.

Let $A = \{a_0, .., a_N\}$ be the set of articles to annotate, of cardinality N; let $P = \{p_0, .., p_M\}$ the set of paragraphs within a given article $a_i \in A$ and cardinality M,.

To keep track of the progress at both global and article-level, we define three sets as follows:

- $COMPLETE$, containing articles for which the amount of annotated paragraphs[7] is larger or equal than a fixed threshold th (we set $th = 5$, consistently with the data selection procedure described above);
- $STARTED$, containing articles for which the amount of annotated paragraphs is smaller than the above threshold th;
- $READY$, containing articles that have never been submitted for annotation.

[6] Negative sampling is used to obtain negative examples from an existing dataset; for instance, assume an Information Retrieval scenario in which the task is to select the most relevant paragraph for a given query, given the whole collection of paragraphs available. In this case, being able to select paragraphs which come from the same article, but are not the most relevant (i.e. that contain the answer), can help increase the discriminatory power of the model learnt.

[7] A paragraph is considered as *annotated* if it is associated to more than 3 question/answer pairs.

Before the annotation process starts, all articles a_i within the set A belong to the $READY$ set – which will be then treated as a LIFO queue; the queue is initialized with A's elements after shuffling. As a user connects to the platform for the first time, a random article a_i is picked from $READY$, and its paragraphs are shuffled and loaded in the UI.

Once the user has completed his annotation for the first paragraph, the article a_i is moved from $READY$ to $STARTED$; the profile of the user is updated accordingly (see 7.3). When 5 or more paragraphs are complete with annotations for a given article, the article moves to the $COMPLETE$ status.

7.3 Annotators Profile

Besides credentials and other information, the annotator profile should include:

- *working_article*: the id of the article a user has started to work on;
- *completed_jobs*: the ids of the articles a user has completed.

7.4 Article Orphanage

Further, we set an expiration date exp for a job at e.g. 48 hours. If user u_i has started to work on article a_j and then logged off without finishing, article a_j would be in the $STARTED$ set; if exp elapses without a_j moving to the $COMPLETE$ queue, then a_j is added on top of the $READY$ queue – meaning it could be reassigned to any connecting user.

8 Next steps

The dataset will be released publicly under CC-BY-SA license. The dissemination efforts will include the publication of one or more scientific articles.

References

[1] Pengyuan Liu, Yuning Deng, Chenghao Zhu, and Han Hu. Xcmrc: Evaluating cross-lingual machine reading comprehension. *arXiv preprint arXiv:1908.05416*, 2019.

[2] Yiming Cui, Ting Liu, Li Xiao, Zhipeng Chen, Wentao Ma, Wanxiang Che, Shijin Wang, and Guoping Hu. A span-extraction dataset for chinese machine reading comprehension. *arXiv preprint arXiv:1810.07366*, 2018.

[3] Kyungjae Lee, Kyoungho Yoon, Sunghyun Park, and Seung-won Hwang. Semi-supervised training data generation for multilingual question answering. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan, May 2018. European Languages Resources Association (ELRA).

[4] Seungyoung Lim, Myungji Kim, and Jooyoul Lee. Korquad: Korean qa dataset for machine comprehension. 2018.

[5] Irene M Cramer, Jochen L Leidner, and Dietrich Klakow. Building an evaluation corpus for german question answering by harvesting wikipedia. In *LREC*, pages 1514–1519, 2006.

[6] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 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, 2016.