etalab gouv.fr

reciTAL.

Projet PIAF

PROTOCOLE

etalab gouv.fr

reciTAL.

1 Objectifs

Les applications du Traitement Automatique des Langues (TAL), qui opèrent sur des langages naturels, sont nécessairement dépendantes des spécificités de ces différents langages. Cependant, il existe une asymétrie des données d'entraînement en TAL, qui sont très majoritairement en anglais. L'importance de ce problème a été reconnue récemment par une résolution du Parlement Européen. [1]

La communauté scientifique a produit au fil des ans de nombreuses ressources pour résoudre des tâches que nous désignerons par simplicité comme étant tâches "bas-niveau" : l'étiquetage morpho-syntaxique (POS-tagging en anglais) ou l'analyse syntaxique, par exemple. Ces tâches bas-niveau possèdent des données d'entraînement en plusieurs langues ou des algorithmes agnostiques à la langue utilisée. Cependant, pour les tâches haut-niveau, qui permettent le développement de produits à valeur ajoutée tels que les systèmes de Questions/Réponses (QR) et les agents conversationnels, les approches à l'état de l'art basées sur de l'apprentissage profond nécessitent des quantités massives de données qui sont disponibles aujourd'hui presque exclusivement en anglais. Le champ de la traduction automatique est une des exceptions notables à ce manque de données, où des corpus multilingues parallèles ont été créés notamment à partir de ressources de l'Union Européenne, tout comme les algorithmes de modélisation de langue, qui peuvent être utilisés de manière semi-supervisée et ainsi n'utiliser que du texte non annoté dans la langue cible.

Un écart important de performance est observé lorsque que l'on crée un système de QR en français entraîné sur des données en anglais traduites automatiquement vers le français. Nous pouvons émettre l'hypothèse que cette chute face aux systèmes de QR anglophones est au moins partiellement causée par les limites des modèles de traduction automatique utilisés. La question du transfert des performances de modèles à l'état de l'art anglophones à d'autres langues que l'anglais reste cependant ouverte. Bien que les méthodes basées sur la traduction, ou les modèles multilingues, peuvent obtenir des résultats raisonnables, des publications récentes [1] sur la langue chinoise (sans doute la langue avec le plus de ressources d'entraînement disponibles après l'anglais) indiquent qu'il existe une marge d'amélioration importante pour les modèles non anglophones. Pour permettre à la communauté scientifique de s'attaquer à ces problèmes, il faut obtenir des données naturelles (c'est-à-dire écrites par des locuteurs natifs et sans biais de traduction), comparables aux données anglophones existantes pour la tâche et la langue choisies. Des efforts en ce sens se sont concentrés sur le chinois [2], le coréen [3, 4], et, dans une moindre mesure, l'allemand [5].

Dans le but de réduire ce manque de données disponibles en français, nous présentons dans ce document le protocole que nous suivons pour collecter un ensemble de données francophones similaire à l'ensemble de données de QR très populaire nommé "Stanford Question Answering Dataset" (SQuAD) [6].

2 Le déroulement de la collecte

Nous ne souhaitons pas faire appel à des "plateformes de travail du clic" (comme par exemple Amazon Mechanical Turk). La collecte de données se déroulera ainsi au cours de plusieurs événements. Outre la valeur pour la recherche et l'industrie des données qui seront collectées, ce projet vise à impliquer la communauté francophone tout en contribuant à sensibiliser les administrations publiques aux apports de l'intelligence artificielle (IA). Dans le même temps, cela permettra de comprendre comment les modèles actuels sont entraînés et d'attirer l'attention sur certains biais qu'ils contiennent. Ce sera également une opportunité pour mesurer l'attractivité et l'efficacité d'un effort de collaboration publique. Nous nous appuyons sur la version française de Wikipédia : nous sélectionnons les articles pertinents, les segmentons en paragraphes plus petits et nous collectons des ensembles de paires de question/réponse correspondant à ces paragraphes. Deux principaux modes de collecte sont envisagés : le premier, *certifié*, dans lequel la participation à la collecte est encadrée (formations, évènements en présentiel, partenariats, etc.) ; le second, *contributeur*, sera diffusé à une plus large participation. Ces deux modes sont ouverts à la participation en ligne.

1. Résolution du Parlement européen du 11 septembre 2018 sur légalité des langues à lère numérique. `http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P8-TA-2018-0332+0+DOC+XML+V0//FR`

L'ensemble de données produit comprendra trois sous-ensembles séparés, à savoir :
— *entraînement*, représentant 80% des articles sources ;
— *validation*, représentant 10% des articles sources ;
— *test*, représentant 10% des articles sources.

Nous nous référons à l'union des ensembles *validation* et *test* en tant que données de *développement*.

Comme le mode *certifié* est par définition plus contrôlé et que la qualité des données devrait y être supérieure, nous l'utiliserons pour la collecte de données de *développement*. Ce mode pourra également être utilisé pour la collecte de l'ensemble *d'entraînement*. En revanche, le mode *contributeur* ne sera utilisé que pour les données d'entraînement. En effet, les données de développement (qui constituent 20 pour cent des données à récolter) doivent être de haute qualité [2]. Pour l'ensemble d'*entraînement*, la présence de données de qualité inférieure ne nuit pas à la construction de modèles efficaces (et contribue possiblement à rendre les modèles plus robustes, par exemple aux erreurs typographiques).

3 Évaluation continue

Il est important d'expliquer aux contributeurs que les questions ne doivent pas être trop facile à répondre. Dans l'interface de collecte de SQuAD, il était précisé que l'annotateur devait tenter de reformuler les mots et expressions du paragraphe. Dans notre interface d'annotation, nous donnerons des exemples expliqués de bonnes et mauvaises questions. Nous détaillons dans cette section les procédures d'évaluation de la qualité des questions.

Pour vérifier la qualité des données collectées, des évaluations manuelles et automatiques seront effectuées tout au long du processus de collecte de données. Celles-ci seront effectuées en continu, à mesure que l'ensemble de *développement* collecté augmente. Nous illustrons ci-dessous le procédés d'évaluation manuelle et automatique de qualité des questions par des exemples pour plus de clarté.

3.1 Manuelle

Nous sous-échantillonnons les données et évaluons les triplets ($\{Paragraphe, Question, Reponse\}$) en fonction du raisonnement requis :

— *variation lexicale* : moins la question et le paragraphe ont des mots en communs, plus cette mesure sera élevée. Par exemple, pour la phrase, "*À l'époque gauloise, le site de Nantes appartient au territoire des Namnètes.*", la question "*À qui appartient le site de Nantes à l'époque gauloise ?*" présente peu de variablité lexicale, contrairement à "*Qui sont les habitants de Nantes avant l'arrivée des Romains ?*"
— *variation syntaxique* : nous détaillons cette mesure, qui peut aussi être calculée automatiquement, dans la section suivante.
— *raisonnement en phrases multiples* : il est nécessaire de combiner les informations contenues par plusieurs phrases pour répondre. Par exemple, soit la séquence de phrases suivante : "*À l'époque gauloise, le site de Nantes appartient au territoire des Namnètes, [...]. Les Romains latinisent son nom gaulois en Condevincum.*". Répondre "*Nantes*" à la question "*Quel est le nom actuel de la ville romaine de Condevincum ?*" nécessite d'utiliser deux phrases pour répondre.
— *ambiguïté* : la question n'est pas claire, et/ou a plusieurs réponses possibles dans le texte. Soit le paragraphe ci-dessous : "*Après la chute de l'Empire romain d'Occident en 476, la cité de Nantes passe rapidement sous le contrôle du royaume franc de Clovis malgré la résistance des Armoricains et des soldats bretons installés par l'Empire romain depuis 280 environ.*" La question "*Qui occupait Nantes au Ve siècle ?*' n'amène pas de réponse unique évidente.

Pour ces évaluations, nous ciblerons un très petit sous-échantillon de l'ensemble *développement*. Pour le jeu de données SQuAD d'origine, cela représente un total de 192 questions, soit 4 par article.

2. Par rapport aux évaluations de qualité en section 3.1.

3.2 Automatique

Nous utiliserons des scripts pour évaluer automatiquement les divergences syntaxiques entre les questions et les réponses correspondantes. Cette procédure sera appliquée séparément à chaque groupe de données. La divergence syntaxique est définie comme la distance d'édition entre les chemins de dépendance non-lexicalisés dans la question et la phrase contenant la réponse. Cette distance fournit une approximation pour comprendre la complexité des questions posées, cest-à-dire la quantité de raisonnement nécessaire pour y répondre. Par exemple, supposons une question "*Qui est allé à Wittenberg pour écouter Luther ?*" et un texte correspondant "*Des étudiants se sont pressés à Wittenberg pour écouter Luther.*". La comparaison des chemins de dépendance entre les deux phrases montrent qu'elles sont égales ; autrement dit, la structure des deux phrases est la même et la quantité de raisonnement nécessaire pour répondre "*Qui*" avec "*Luther*" est minimale. À linverse, supposons que la question "*Quel impact le mouvement de massification de l'enseignement secondaire a-t-il eu sur la présence de travailleurs qualifiés ?*" avec le texte correspondant "*Au cours du mouvement de massification de l'enseignement secondaire de 1910 à 1940, il y a eu une augmentation du nombre de travailleurs qualifiés*", la mesure de divergence syntaxique indique qu'un raisonnement plus complexe est nécessaire pour fournir une réponse.

En outre, nous mesurerons progressivement les performances des systèmes de QR multilingues les plus performants, sous différentes configurations expérimentales. Les résultats de cette dernière étape dévaluation déclencheront éventuellement le passage à une étape ultérieure de la procédure de collecte de données.

4 Collecte de données

La collecte de données commence par cibler un ensemble de données comparable à SQuADv1.1, pour lequel la grande majorité des questions devrait avoir une réponse exacte dans le paragraphe correspondant [3]. La priorité sera de collecter des données d'évaluation de qualité, à hauteur de 20 000 environ. Nous visons ensuite de collecter autour de 80 000 données d'entraînement.

Nous envisageons plusieurs étapes de la collecte de données, que nous détaillons ci-dessous. En fonction de l'évaluation continue décrite dans la section 3.2, nous déclencherons éventuellement le passage aux étapes suivantes de la collecte de données, conçues pour augmenter la difficulté de l'ensemble de données.

4.1 Collecte de base (SQuADv1.1)

Ce mode sert de base à tous les modes de collecte envisagés et vise à collecter des données comparables à SQuADv1.1. À cette fin, deux versions de l'interface utilisateur sont requises et décrites dans les sections 4.1.3 et 4.1.4. Nous supposons que chaque participant est associé à un identifiant unique pouvant les lier à un statut *contributeur certifié* ou *contributeur*.

4.1.1 Sélection des données sources

Conformément à SQuAD, les articles de Wikipédia à utiliser comme source pour la tâche de QR seront sélectionnés selon la méthodologie suivante, avec les paramètres expliqués dans la section 7.1. Tout dabord, le score de centralité PageRank [4] sera calculé [5] pour chaque page de l'espace principal de Wikipédia en français, afin de classer les articles et de sélectionner les plus pertinents. Enfin, nous sélectionnerons chacun des paragraphes, en supprimant les images, les figures, les tableaux et les paragraphes de moins de 500 caractères. Comme mentionné ci-dessus, les articles seront répartis au hasard en 80-10-10, pour entraînement-validation-test respectivement.

3. pour SQuADv1.1, le nombre de questions marquées comme étant sans réponse dans le *développement* ne représente que 2,6%
4. `https://en.wikipedia.org/wiki/PageRank`
5. Pour ce calcul nous utiliserons le code publié par le projet Nayuki, `https://www.nayuki.io/page/computing-wikipedias-internal-pageranks`, avec des modifications pour l'adapter au français

4.1.2 Affichage des instructions

Après s'être connecté, un exemple de paragraphe, des exemples de bonnes et de mauvaises questions réponses concernant ce paragraphe, ainsi qu'une explication de la raison pour laquelle elles ont été classées comme telles sont montrés à l'utilisateur.

4.1.3 Collecte des données d'entraînement

Les utilisateurs sont invités à choisir un paragraphe parmi ceux pour lesquels aucune paire question/réponse n'est encore associée. Sur la figure 1, nous présentons une capture d'écran de l'interface utilisateur d'origine utilisée pour la collecte des données dans ce mode. La fonctionnalité copier/coller doit être désactivée : un utilisateur ne doit pas pouvoir copier du texte de la zone de paragraphe dans les zones de questions. Lorsque le bouton "sélectionner une réponse" est cliqué, l'utilisateur peut sélectionner/mettre en surbrillance une zone dans le paragraphe, qui sera enregistrée en tant que réponse à la question écrite correspondante. L'annotateur peut choisir de soumettre le paragraphe actuel avec les paires question-réponse générées ou de continuer à générer des questions. L'interface utilisateur doit amener l'annotateur à générer cinq paires de questions-réponses par paragraphe avant la soumission.

4.1.4 Collecte des données d'évaluation

Afin de rendre les données dévaluation plus robustes et dobtenir des mesures de la performance humaine de la tâche, des réponses supplémentaires sont collectées pour toutes les questions des ensembles *validation* et *test* (l'ensemble développement).

À cette fin, un utilisateur est d'abord informé des instructions concernant la tâche. Ensuite, ils ne voient qu'un paragraphe et une question précédemment écrite par un autre utilisateur. Sa tâche consiste alors simplement à sélectionner la réponse à la question si elle existe, ou à marquer la question comme *impossible* sinon (dans SQuADv1.1, cela était supposé être le cas lorsque l'utilisateur ne sélectionnait pas de réponse à ce stade).

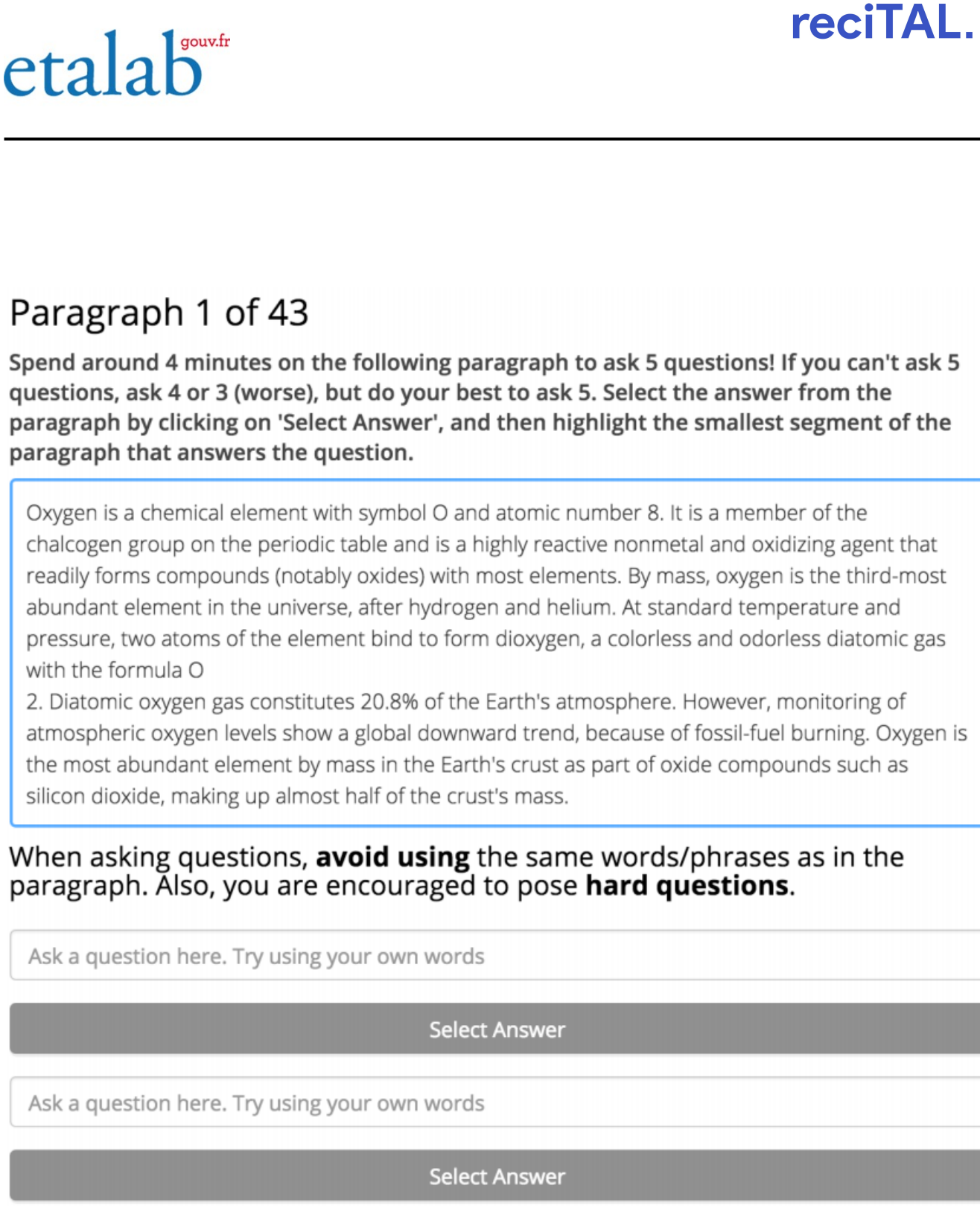

FIGURE 1 – Interface utilisateur pour la collecte de SQuADv1.1

4.2 Format des données

Tous les jeux de données partageront une structure commune, compatible avec le format SQuAD :

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

Comme indiqué ci-dessus, la clé `data` correspond à une liste d'articles de Wikipédia, structurés comme suit :

```
{'title': '' # type: string -- titre de la page wikipedia
 'paragraphs': [] # type: list -- liste des paragraphes dans la page wikipedia
}
```

La structure des éléments de la liste `paragraphs` est cohérente parmi les ensembles entraînement, développement et test.

Le format de sortie pour les paragraphes de données d'entraînement est le suivant :

```
{'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'
        }]}
```

Pour chaque question, les paires de QR d'entraînement ne doivent avoir qu'une seule réponse, tandis que celles d'évaluation peuvent avoir plusieurs réponses.

5 Changement de protocole

Après une évaluation automatique (voir section 3.2) nous devrions pouvoir passer à tout moment à un protocole de collecte légèrement différent, afin dobtenir des questions de plus en plus difficiles. Nous déclencherons un basculement au cas où l'évaluation automatique de l'ensemble de développement suggère que les résultats sont comparables à ceux de SQuAD. Les questions/réponses générées dans un mode spécifique doivent être marquées comme tels via un marqueur spécifique.

5.1 Questions impossibles à répondre

Dans ce mode, nous allons collecter des paires de QR *plausibles mais impossibles à répondre* (aucune réponse ne se situe dans le paragraphe), en ciblant au maximum 50% des données déjà collectées. Ceci est cohérent avec SQuAD version 2.0 ; l'interface utilisateur utilisée pour SQuADv2.0 est illustrée à la figure 2. Outre les modifications apportées aux instructions, la seule différence, en termes dinterface utilisateur, par rapport à celle indiquée dans la figure 1 réside dans laffichage de "Questions pour inspiration'' : ce sont les questions *avec réponse correcte* déjà collectées à l'étape précédente de la collecte.

En termes de format, ce paramètre nécessite lajout de deux nouveaux champs *plausible_answers* (réponses plausibles, mais fausses) et *is_impossible* (est impossible) aux éléments de la liste *qas*.

```
{'plausible_answers': [{'text': 'Normandie', 'answer_start': 137}],
 'question': 'La France est une région de quel pays ?',
 'id': '5ad39d53604f3c001a3fe8d2',
 'answers': [],
 'is_impossible': True}
```

5.2 Mode adversaire

Afin d'inciter les contributeurs à poser des questions difficiles, nous utiliserons un modèle de QR pré-entraîné. Les utilisateurs ne seront invités à soumettre une question que si le système ne parvient pas à y répondre correctement. Dès qu'un utilisateur pose une question, le service de QR est appelé. La réponse renvoyée sera mise en évidence dans le paragraphe et il sera demandé à l'utilisateur de soumettre sa paire de question/réponse uniquement si le service de QR s'est trompé.

Comme base de référence, nous utiliserons le modèle BERT multilingue pré-entraîné sur les données d'entraînement en anglais de SQuAD. Ce modèle sera évalué à la fois sur les ensembles de développement anglais et français pour mesurer et comparer les performances. Cela permettra également de mieux mesurer limpact des techniques daugmentation des données (par exemple, la traduction automatique des données d'entraînement).

6 Estimations temporelles

Dans le tableau 1, nous rapportons les estimations temporelles nécessaires pour rassembler le corpus dans le cadre de scénarios basés sur SQuAD (les données des articles SQuAD sont utilisées).

En supposant 20 volontaires par session, avec une session comprenant 4 heures de temps dannotation effectif, nous pouvons espérer collecter une quantité de données dévaluation comparable à SQuAD (20k) en 4 ou 5 sessions (pour $n\_req = 2$, où n_req est le nombre minimal de réponses additionnelles nécessaires), soit un total d'entre 320 et 400 heures.

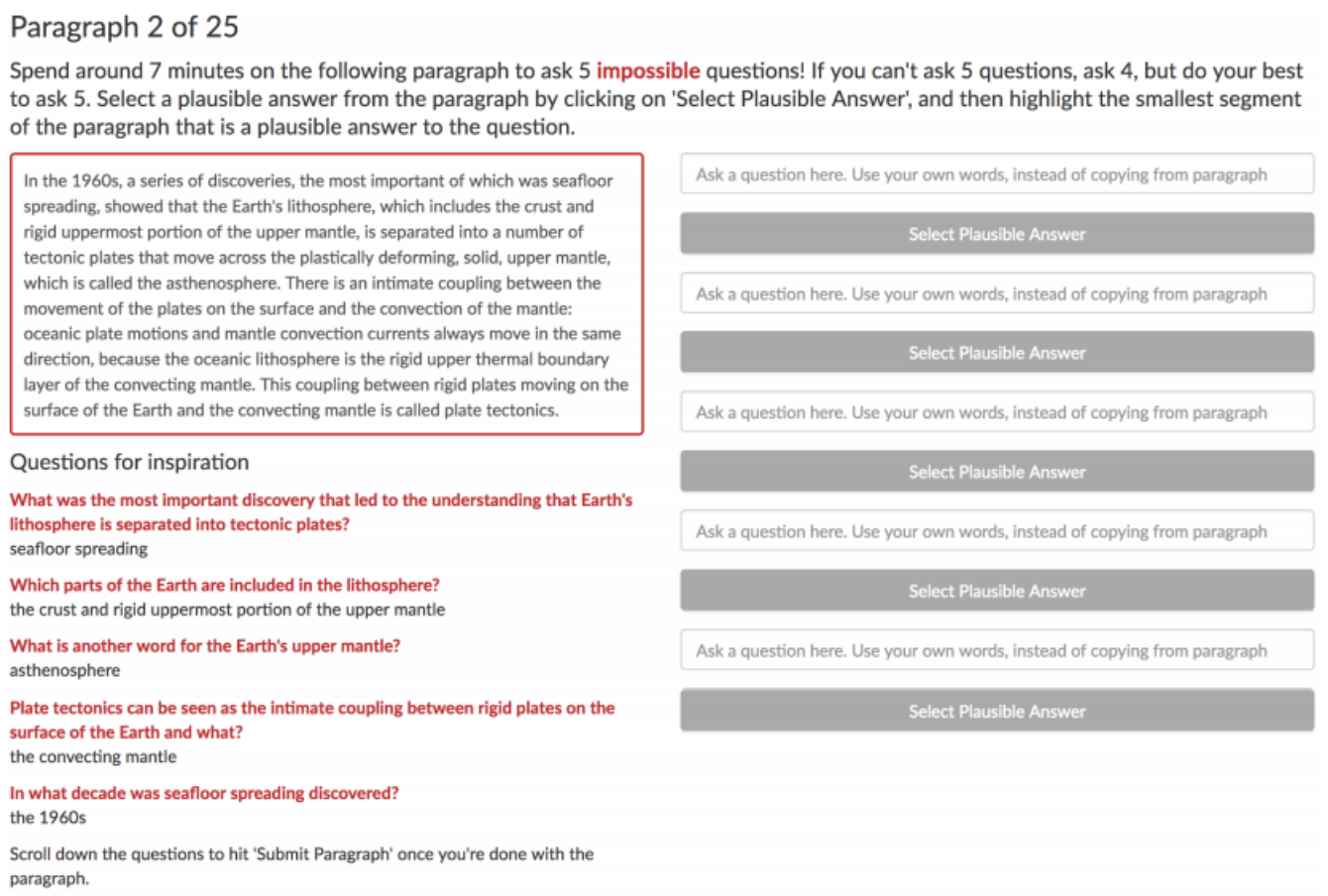

FIGURE 2 – UI pour la collecte de questions *impossibles à répondre* sur SQuADv2.0

mode	*paire QR/minute*	*participant/h pour 10k paires*	*paires par session*
base (entraînement)	1.25	133.3	6000
base (additionel/dev)	2.5	66.6 (x n_req)	$12000/n\_req$
questions impossibles	1.4	233.3	6720

TABLE 1 – Temps estimé pour chaque mode. n_req est le nombre minimal de réponses additionnelles nécessaires (pour SQuADv1.1, $n\_req = 2$).

7 Déroulement

7.1 Sélection des articles Wikipédia

Dans SQuAD, les auteurs ont utilisé le top 10 000 du Wikipédia anglophone, c'est-à-dire les articles les mieux classés selon PageRank. Néanmoins, en appliquant le même seuil à la version française, nous avons constaté des différences significatives en termes de propriétés structurelles entre les Wikipédia français et anglais. Par exemple, nous avons observé la présence massive de pages faisant référence à des *années* sur la version française, une caractéristique que le sous-échantillon de top 10 000 de Wikipédia anglais ne semble pas avoir. Après l'inspection manuelle, les pratiques d'édition semblent différer entre les communautés française et anglaise : si cette dernière ne lie pas toutes les mentions d'*année* à la page consacrée à cette année, les éditeurs de Wikipédia français semblent le faire systématiquement, un fait qui augmente le score de PageRank de ces articles et explique donc leur présence très importante dans le sous-échantillon français.

De plus, étant donné la présence importante darticles de Wikipédia non exploitables pour nos objectifs (brouillons, pages dhomonymie, etc.), nous fonctionnons comme suit :

— rassembler les 25 000 articles les plus populaires en termes de PageRank ;
— éliminer les articles inexploitables ;
— définir une limite de caractères min-max sur la longueur du paragraphe ($min = 500; max = 1000$) ;
— filtrer les articles avec moins de 5 paragraphes ;
— appliquer la procédure de fractionnement entraînement/validation/test.

En comparant au SQuAD anglais, nous espérons donc obtenir des données de QR annotées sur plus d'articles, avec moins de paragraphes par article et une longueur comparable. Ce compromis semble raisonnable : le principal inconvénient est la disponibilité de moins de paragraphes par article, mais comme ce facteur est surtout pertinent pour l'échantillonnage négatif [6], un ratio 5 : 1 semble acceptable (c'est-à-dire cinq paragraphes par article).

Pour chaque article, un identifiant unique est calculé (par exemple. en prenant le hash du titre). De plus, chaque paragraphe se voit attribuer un identifiant unique composé de l'identifiant de l'article auquel il appartient, ainsi qu'un numéro séquentiel correspondant à sa position dans l'article.

7.2 Algorithme de sélection des paragraphes

L'algorithme permettant de sélectionner le paragraphe à afficher pour un utilisateur est défini comme suit.

Soit $A = \{a_0, .., a_N\}$ l'ensemble des articles à annoter, de cardinalité N ; soit $P = \{p_0, .., p_M\}$ l'ensemble des paragraphes dans un article donné $a_i \in A$, de cardinalité M.

Pour suivre les progrès au niveau global et au niveau des articles, nous définissons trois ensembles comme suit :

— $COMPLETE$ (complet), contenant des articles pour lesquels la quantité de paragraphes annotés [7] est supérieur ou égal à un seuil fixe th (nous avons défini $th = 5$, conformément à la procédure de sélection de données décrite ci-dessus) ;
— $STARTED$ (en cours), contenant des articles pour lesquels le nombre de paragraphes annotés est inférieur au seuil ci-dessus th ;
— $READY$ (prêt), contenant des articles pour lesquels aucun paragraphe n'a été annoté.

Ainsi, avant que le processus d'annotation ne commence, tous les articles de l'ensemble A appartiennent à l'ensemble $READY$ qui sera traité comme une file d'attente LIFO (dernier arrivé, premier sorti) ; la file d'attente est initialisée avec les éléments de A après avoir été mélangé. Lorsqu'un utilisateur se connecte à la plate-forme

6. L'échantillonnage négatif est utilisé pour obtenir des exemples négatifs à partir d'un jeu de données existant. Par exemple, supposons un scénario d'extraction d'informations dans lequel la tâche consiste à sélectionner le paragraphe le plus pertinent pour une requête donnée, compte tenu de l'ensemble des paragraphes disponibles. Dans ce cas, le fait de pouvoir sélectionner des paragraphes qui proviennent du même article, mais qui ne sont pas les plus pertinents (c'est-à-dire contenant la réponse), peut aider à augmenter le pouvoir discriminatoire du modèle appris.

7. Un paragraphe est considéré comme *annoté* s'il est associé à plus de 3 paires question / réponse.

pour la première fois, un article aléatoire a_i est choisi dans $READY$ et ses paragraphes sont mélangés et chargés dans l'interface utilisateur.

Une fois que l'utilisateur a terminé l'annotation pour le premier paragraphe, l'article a_i est déplacé de $READY$ à $STARTED$; le profil de l'utilisateur est mis à jour en conséquence (voir section 7.3). Lorsque 5 ($th = 5$) paragraphes ou plus sont complets avec des annotations pour un article donné, l'article passe au statut $COMPLETE$.

7.3 Profil des annotateurs

Outre les informations d'identification et d'autres informations, le profil d'annotateur doit inclure :
— *working_article* : l'identifiant de l'article sur lequel un utilisateur a commencé à travailler ;
— *completed_jobs* : les identifiants des articles qu'un utilisateur a complétés.

7.4 Article orphelin

En outre, nous fixons une date d'expiration *exp* pour un travail, par exemple 48 heures. Si l'utilisateur u_i avait commencé à travailler sur l'article a_j, puis s'était déconnecté sans avoir terminé, l'article a_j est déplacé dans l'ensemble $STARTED$. Si *exp* s'écoule sans que a_j ne soit déplacé vers la file d'attente $COMPLETE$, alors a_j est ajouté au-dessus de la file d'attente $READY$, ce qui signifie qu'il peut être réaffecté à tout utilisateur connecté.

8 Prochaines étapes

Le jeu de données produit sera partagé sous licence CC-BY-SA. Le travail qui en sortira sera présenté à la communauté académique sous la forme dune publication scientifique.

Références

[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.