CALOR-QUEST : un corpus d'entraînement et d'évaluation pour la compréhension automatique de textes

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

RÉSUMÉ

La compréhension automatique de texte est une tâche faisant partie de la famille des systèmes de *Question/Réponse* où les questions ne sont pas à portée générale mais sont liées à un document particulier. Récemment de très grand corpus (SQuAD, MS MARCO) contenant des triplets (document, question, réponse) ont été mis à la disposition de la communauté scientifique afin de développer des méthodes supervisées à base de réseaux de neurones profonds en obtenant des résultats prometteurs. Ces méthodes sont cependant très gourmandes en données d'apprentissage, données qui n'existent pour le moment que pour la langue anglaise. Le but de cette étude est de permettre le développement de telles ressources pour d'autres langues à moindre coût en proposant une méthode générant de manière semi-automatique des questions à partir d'une analyse sémantique d'un grand corpus. La collecte de questions *naturelle* est réduite à un ensemble de validation/test. L'application de cette méthode sur le corpus CALOR-Frame a permis de développer la ressource CALOR-QUEST présentée dans cet article.

ABSTRACT

Machine reading comprehension 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, MS MARCO) containing triplets (document, question, answer) were made available to the scientific community to develop supervised methods based on deep neural networks with promising results. These methods need very large training corpus to be efficient, however such kind of data only exists for English at the moment. The aim of this study is the development of such resources for other languages by proposing to generate in a semi-automatic way questions from the semantic Frame analysis of large corpora. The collect of *natural questions* is reduced to a validation/test set. We applied this method on the French CALOR-Frame corpus to develop the CALOR-QUEST resource presented in this paper.

MOTS-CLÉS : Compréhension automatique de texte, Question Réponse, Analyse en cadre sémantique, Génération de questions.

KEYWORDS: Machine reading comprehension, Question Answering, Semantic Frame analysis, Question generation.

1 Introduction

La compréhension automatique de texte (*Machine Reading Comprehension*) consiste pour une machine à répondre à des questions portant sur un document écrit, chaque réponse se présentant sous la forme d'un passage du document. Il s'agit de la version automatique de la tâche, de *compréhension écrite*, qui permet de vérifier les capacités des élèves à lire un texte et à le comprendre. Du point de vue du Traitement Automatique des Langues il s'agit d'un cas particulier de la tâche de *Question/Réponse* où les questions ne sont pas de portée générale mais sont, au contraire, liées à un document particulier.

Cette tâche a reçu récemment une attention particulière avec la mise à disposition de deux très grands corpus de textes enrichis de questions et réponses : le corpus **SQuAD** (Rajpurkar *et al.*, 2016) et le corpus **MS MARCO** (Nguyen *et al.*, 2016) contenant chacun plus de 100k triplets *(corpus, question, réponse)* où chaque question a été produite par un humain, soit via crowd-sourcing, soit via de vraies requêtes de moteurs de recherche. Ces corpus ont permis le développement de nombreuses approches basées sur de l'apprentissage supervisé, principalement avec des réseaux de neurones profonds, tel que (Wang & Jiang, 2016) ou (Seo *et al.*, 2016), avec une amélioration significative des résultats par rapport aux méthodes fondées sur des analyses linguistiques ou sur des méthodes d'appariement entre questions et texte contenant les réponses (Hermann *et al.*, 2015).

Ces ressources sont disponibles en langue anglaise mais pour d'autres langues, telles que le français, il n'existe pas de corpus comparable et l'effort nécessaire pour collecter une telle quantité de données est très important, limitant l'emploi de ces méthodes à d'autres langues ou à d'autres cadres applicatifs. Pour répondre à ce problème, l'étude présentée dans ce papier vise à créer, de manière partiellement automatique, un corpus permettant d'entraîner de tels systèmes de manière supervisée sans avoir à collecter manuellement une très grande collection de données. La méthode proposée consiste à utiliser une analyse en cadre sémantique de type *FrameNet* sur de grands corpus de texte, puis à générer automatiquement des questions à partir des analyses obtenues. Ces questions automatiques peuvent servir à entraîner des méthodes de compréhension de textes qui seront évaluées sur des questions *naturelles* posées par des évaluateurs humains sur les mêmes documents.

2 Travaux similaires

Plusieurs corpus pour la tâche de Compréhension de Texte par la Machine (Machine Reading Comprehension) sont disponibles sous la forme de paires de question/extrait de texte pouvant répondre à la question. Si le corpus **SQuAD** a généré de nombreuses contributions, le corpus **MS-MARCO** produit à partir de requêtes sur le moteur de recherche Bing en est également un exemple récent. On peut se référer à l'article décrivant ce dernier (Nguyen *et al.*, 2016) pour une revue détaillée d'autres corpus en langue anglaise issus de différentes sources comme **NewsQA** (Trischler *et al.*, 2016), **SearchQA** (Dunn *et al.*, 2017) proposant des questions du jeu Jeopardy appariées à des extraits de textes issus de requêtes Google, **NarrativeQA** (Kočiský *et al.*, 2018) construit à partir de résumés de films et de livres. On y trouve également différents types de questions comme des questions à choix multiples dans le corpus **ARC** (Clark *et al.*, 2018) ou des questions insérées sous forme de textes à trous comme dans le corpus **ReCoRD** (Zhang *et al.*, 2018) qui vise à tester l'influence de la connaissance du sens commun. On y trouve également quelques références à des corpus en langue chinoise mais à notre connaissance il n'existe pas de telles ressources en langue française.

Nous cherchons dans cette étude à construire un tel corpus pour le français sans pour autant devoir mettre en oeuvre un processus d'annotation trop complexe. Nous proposons une méthode semi-

automatique partant d'un texte et d'une représentation sémantique de ce texte pour produire les questions associées. La génération de questions à partir d'un texte est une problématique ancienne ayant fait l'objet de nombreux travaux, notamment à l'occasion de campagnes d'évaluation (Boyer & Piwek, 2010). Deux grandes familles de méthodes ont été explorées, que ce soit grâce à des patrons construits à partir de l'analyse syntaxique d'une phrase ou à partir d'une analyse sémantique. Les progrès récents dans ces deux disciplines ont permis de nouvelles avancées en génération de question (Mazidi & Nielsen, 2014). Récemment (Pillai *et al.*, 2018) et (Flor & Riordan, 2018) par exemple proposent de générer des questions factuelles à partir d'une analyse en rôles sémantiques de type PropBank. En revanche ces travaux se situent bien souvent dans un contexte applicatif différent du nôtre, à savoir la production de question pour l'apprentissage d'une langue ou la génération de quizz dans un processus pédagogique. Dans de tels contextes, la lisibilité et la grammaticalité des questions obtenues est primordiale et les questions sont évaluées par des tests subjectifs ou des métriques de type *Bleu* ou *Meteor*. Au delà des approches par règles, des travaux récents envisagent la génération de question comme une tâche d'apprentissage à part entière où la question est générée directement grâce à un réseau de neurones à partir du texte et d'un conditionnement par la réponse (Dong *et al.*, 2018) (Yuan *et al.*, 2017) voire même en envisageant conjointement les tâches de génération de question et de réponse (Wang *et al.*, 2017). Nous proposons une approche de construction de questions à l'aide de patrons reposant sur une analyse sémantique de type FrameNet, nous permettant d'obtenir un corpus utilisable pour étudier la compréhension automatique de texte sur le français.

3 Génération semi-supervisée d'un corpus de questions

Le corpus CALOR-Frame est composé de 4 sous-corpus issus de 3 sources encyclopédiques : Wikipédia (WP), Vikidia (V) et ClioTexte (CT). 3 thématiques sont représentées : la première guerre mondiale (1GM), l'archéologie (arch) et l'antiquité (antiq). La variété des sources permet de couvrir différents registres de langage allant du document historique pour ClioTexte (discours, déclarations) aux articles pour enfants dans Vikidia. Ce corpus a été annoté manuellement en cadre sémantique selon le schéma d'annotation *Berkeley FrameNet* (Baker *et al.*, 1998) avec un ensemble de 54 *Frames* décrit dans (Béchet *et al.*, 2017). Les cadres sémantiques ou *Frames* décrivent des situations prototypiques (*décider, perdre, attaquer, vaincre, etc.*). L'annotation consiste d'abord à identifier le mot déclencheur de la Frame, appelé *Lexical Unit (LU)*, puis les actants et circonstants appelés *Frame Elements (FE)*. Le nombre de déclencheurs différents représentés dans le corpus correspond à 145 lemmes (70 noms et 75 verbes). Une même séquence de mots peut correspondre à plusieurs *FEs* différents si une phrase comporte plusieurs *Frames*. Un exemple est donné figure 1 pour une phrase annotée avec les deux *Frames*, `Losing` déclenchée par le mot `perdu` et `Attack` déclenchée par `attaques`.

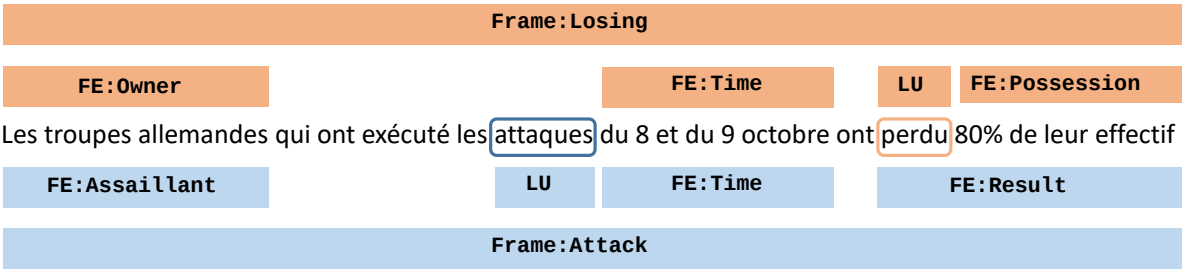

FIGURE 1 – Exemple de phrase annotée avec des cadres sémantiques provenant de FrameNet

Ce type d'annotations est particulièrement intéressant pour l'obtention d'un corpus de *ques-*

tions/réponses pour le développement de méthodes de compréhension automatique de texte. En effet, en sélectionnant une *Frame* F et un *Frame Element* E dans une phrase, on peut produire automatiquement des questions dont la réponse est E. La Frame ainsi que les autres *FE* présents dans la phrase vont constituer le contexte C de ces questions que l'on va noter sous forme de triplets : (F, E, C). En faisant varier F, E et C pour une même phrase, on peut obtenir un ensemble de questions pour lesquelles on dispose des réponses avec leur classe sémantique (le type du *FE*).

Par exemple, sur la phrase de l'exemple 1, nous aurons 18 combinaisons possibles (F, E, C) : `(Losing, Owner, {Time, Possession})`, `(Losing, Owner, {Possession})`, `(Losing, Owner, {Time})`, `(Losing, Time, {Owner, Possession})` ...

A chaque triplet (F, E, C) peut correspondre de nombreuses questions $Q \in$ *Questions*(F, E, C). Par exemple pour la première combinaison, on pourrait avoir : *Qui a perdu 80% de ses effectifs du 8 au 9 octobre ?*, ou encore *Quelles troupes ont été décimées à 80% dans les attaques du 8 et colnou9 octobre ?*. Ces deux questions ont pour réponse le même segment de texte (`les troupes allemandes`), mais ont des formes très différentes : la première est très proche de la phrase originale et pourrait être obtenue par une simple réorganisation de celle-ci sans modification d'ordre lexical, la deuxième en revanche suppose une réécriture complète avec ajout éventuel de nouveaux termes (*décimées*). Ces deux types de questions ont été produits à partir de l'annotation en Frame du corpus CALOR-Frame : une méthode automatique a permis de générer des questions Q_A à partir des combinaisons (F, E, C), et une collecte manuelle sur un sous-ensemble du corpus a permis d'obtenir des questions *naturelles* notées Q_N dans un environnement sémantique contrôlé. Ces deux méthodes sont présentées dans les paragraphes suivants. Dans une première approche, nous nous appuyons sur l'annotation en Frames manuelle du corpus CALOR-Frame pour valider les principes généraux de la méthodologie. Des travaux sont en cours pour valider l'approche à partir de corpora annotés automatiquement par un système d'étiquetage en Frames.

Production de questions à partir de patrons

La production automatique de questions à partir de cadres sémantiques repose sur des patrons de questions présentés dans la table 1. Un patron est rattaché à un type de Frame F. Les mots précédés du symbole $ dans les patrons de F correspondent aux types de FE de F. Les éléments entre crochets sont optionnels et les autres sont obligatoires. De ce fait, un patron peut générer plusieurs questions, en fonction des éléments optionnels choisis.

Activity_start	spécifique	Qu'est-ce qui a commencé $Time [$Place] [pour $Purpose] ? ⟶ $*Activity* .
	générique	quoi commencer [$Agent] [$Circumstances] [$Co-timed_event] [$Containing_event] [$Event_description] [$Explanation] [$Manner] [$Means] [$Place] [$Purpose] [$Time] ? ⟶ $*Activity* .
Leadership	spécifique	Quand est-ce que $Leader a dirigé $Governed [$Place] ? ⟶ $*Time* .
	générique	quand diriger [$Leader] [$Governed] [$Place] [$Role] [$Duration] [$Activity] ? ⟶ $*Time* .

TABLE 1 – Exemples de patrons génériques et spécifiques pour les Frames Activity_start et Leadership

Deux type de patrons ont été développés : des patrons génériques à toutes les Frames, se contentant d'introduire la question par un pronom interrogatif correspondant au type de E puis énumérant toutes

les combinaisons de FE du contexte C ; des patrons spécifiques à chaque Frame, spécifiant les FE obligatoires et facultatifs, ne permettant pas toutes les combinaisons de FE possibles afin de garantir une apparence *naturelle* aux questions générées. Les patrons génériques permettent de générer un très grand nombre de questions couvrant tous les cas possibles, sans se soucier de la justesse syntaxique des phrases générées. Au contraire les patrons spécifiques sont censés générer des phrases plus proches de questions naturelles que pourrait poser un utilisateur.

Collecter des questions naturelles

La collecte des questions naturelles s'est faite auprès d'annotateurs à qui l'on présentait les éléments (F, E, C). La phrase originale n'était pas affichée pour laisser plus de liberté aux annotateurs dans les choix lexicaux effectués pour rédiger les questions. Les annotateurs avaient aussi toute liberté dans le choix des éléments du contexte C qu'ils allaient inclure dans leurs questions. L'exemple suivant correspond à la configuration `Q(Hiding_objects, Place, {Agent, Hidden_object, Hiding_place})`.

Frame = `Hiding_objects`

- Contexte
  - **Agent** : `un chef de milice gauloise`
  - **Hidden_object** : `un trésor`
  - **Hiding_place** : `dans sa ferme de Bassing`
- But
  - objet de la question : `Place`
  - réponse : `Moselle`
- Questions collectées :
  - *Dans quelle région un chef de milice gauloise cache-t-il un trésor ?*
  - *Où se trouve la ferme de Bassing, dans laquelle un chef de milice gauloise cache un trésor ?*

Les questions naturelles produites avec ce protocole pour la constitution du corpus CALOR-QUEST ont porté sur un sous ensemble du corpus en termes de documents, mais l'ensemble des cadres sémantiques du corpus CALOR-Frame sont représentés.

Deux ensembles de questions

L'objectif de l'étude n'est pas de valider la qualité intrinsèque des questions générées automatiquement. Si leur construction garantit qu'elles sont valides d'un point de vue sémantique, nous savons qu'elles ne sont pas correctes d'un point de vue syntaxique. Nous ne chercherons donc pas à vérifier si les questions automatiques se retrouvent parmi les questions naturelles. Nous nous intéressons à ce qui relie ces deux types de questions à savoir les occurrences de Frames à partir desquelles elles ont été construites. En effet, une question naturelle et une question automatiques produites à partir d'une même occurrence de Frame et posées sur un même Frame Element ont par construction la même réponse, à savoir ce Frame Element. L'ensemble du corpus avec questions naturelles et questions automatiques sera mis à disposition de la communauté scientifique.

4 Evaluation du corpus *CALOR-QUEST*

Pour cette toute première étude sur le corpus CALOR-QUEST, nous évaluons l'hypothèse de travail selon laquelle la réponse à une question naturelle peut être trouvée en effectuant un appariement avec une question générée automatiquement dont on connaît sans ambiguïté la réponse. Pour chaque

question naturelle Q_N d'un document, nous calculons les similarités textuelles avec toutes les questions automatiques de ce document Q_A, et nous apparions Q_N avec la question automatique $\hat{Q}_A$ donnant la similarité maximale tel que : $\hat{Q}_A = \underset{Q_A}{\operatorname{argmax}}\ \textit{sim}(Q_A, Q_N)$

Nous considérons que cet appariement est correct si $\hat{Q}_A$ a la même réponse que Q_N, et nous le notons $\mathrm{correct}(Q_A, Q_N)$. Cela signifie qu'ils peuvent être générés par le même triplet (F, E, C) tel que : $\{\hat{Q}_A, Q_N\} \in \textit{Questions}(F, E, C)$. La réponse à ces deux questions se trouve ainsi être le support dans le texte pour l'élément E. Le tableau 2 recense les statistiques du corpus utilisé pour l'étude. La dernière colonne représente le nombre moyen d'appariements corrects pour une question naturelle donnée (c'est à dire le nombre moyen de questions automatiques ayant la même réponse), alors que l'avant dernière colonne représente le nombre moyen de candidats à l'appariement.

collection	nb docs	nb questions naturelles (Q_N)	nb moyen Q_N par doc	nb questions automatiques (Q_A)	nb moyen Q_A par doc	nb moyen appariements corrects par Q_N
V_antiq	61	274	4.5	4672	76.6	4.2
WP_arch	96	302	2.4	36259	377.7	4.1
CT_1GM	16	241	15.1	7502	468.9	2.5
WP_1GM	123	319	2.6	50971	414.4	5.1
total	**296**	**1136**	**3.8**	**99404**	**335.8**	**4.1**

TABLE 2 – Description du corpus CALOR-QUEST

4.1 Appariement des questions automatiques et naturelles

Nous étudions comparativement 3 mesures de similarité textuelle $\textit{sim}(Q_A, Q_N)$:

- le cosinus entre sacs de mots pondérés, $cos(Q_N, Q_A)$, qui reste une référence solide dans la famille des mesures basées sur les représentations creuses ;
- le cosinus entre des plongements des questions, $wavg-w2v(Q_N, Q_A)$, ces plongements étant simplement obtenus par la moyenne pondérée des plongements des mots qui les composent ;
- la similarité soft-cosinus $cos_M(Q_N, Q_A)$, qui consiste à introduire dans la formule du cosinus en sacs de mots une matrice de relations entre les mots, calculées à partir des plongements lexicaux (ou *embeddings*) de ces derniers (Charlet & Damnati, 2017).

Plus précisément, cette dernière similarité est donnée par :

$$cos_M(Q_N, Q_A) = \frac{Q_N^t.M.Q_A}{\sqrt{Q_N^t.M.Q_N}\sqrt{Q_A^t.M.Q_A}} \quad (1)$$

où Q_N (resp. Q_A) représente le vecteur de sacs de mots pondérés de la question Q_N (resp. Q_A) et M la matrice dont l'élément $m_{i,j}$ exprime la relation entre les mots i et j. Ici, elle est égale au carré de la similarité cosinus entre les plongements des mots i et j. Les questions sont segmentées en mots, et subissent un prétraitement minimal (suppression des majuscules). Les poids des mots sont les TFIDF où les IDF sont estimés par collection, sur le corpus des questions automatiques.

Dans le cas particulier des questions, les pronoms interrogatifs et les prépositions jouent un rôle très structurant pour analyser finement le sens d'une question. Nous évaluons ainsi l'influence de la conservation des mots creux (liste *NLTK*) ou non dans les représentations de questions, ainsi qu'une variante du soft-cosinus, $cos_{M'}$, qui consiste à ne pas considérer de relations pour les mots creux.

4.2 Evaluation

Les performances d'appariement sont présentées dans le tableau 3. Pour cette évaluation, seul l'appariement de meilleur score est considéré. On constate tout d'abord que les performances sont liées de façon monotone au nombre de Q_A candidates pour chaque Q_N (cf tableau 2). Plus ce nombre est élevé, moins les performances sont bonnes. C'est pourquoi l'on observe les meilleures performances sur le corpus V_antiq dont les documents sont plus courts et le nombre moyen de candidats à l'appariement est moins élévé (76.6 Q_A par document en moyenne) alors que les performances les moins bonnes sont obtenues sur le corpus CT_1GM où le nombre de candidats par document est de 468.9 en moyenne. Le maintien des mots-creux améliore nettement les performances dans le cas du cosinus entre sacs de mots tandis que les performances de cos_M sont dégradées et que leur influence est variable pour $wavg-w2v$. Les mots-creux en commun dans les questions favorisent l'appariement lorsqu'ils sont utilisés de façon stricte, mais la prise en compte des plongements de ces mots (que ce soit à travers la moyenne des plongements ou le soft-cosinus) ajoute du bruit. Les relations sémantiques sont néanmoins bénéfiques pour les mots pleins, ce qui est confirmé par les bonnes performances obtenues avec la version $cos_{M'}$. Une analyse des erreurs a révélé qu'une majorité des erreurs conduisent à apparier une Q_N à une Q_A issue d'une Frame différente (54% des erreurs sur V_antiq, 44% sur WP_arch, 61% sur CT_1GM et même 71% sur WP_1GM). Les erreurs résiduelles concernent majoritairement des erreurs vers une question issue de la même Frame mais portant sur un autre *FE*. Une analyse sémantique préalable des questions ou une prédiction du type de *FE* attendu pourrait améliorer la précision de l'appariement.

collection	cos		$wavg-w2v$		cos_M		$cos_{M'}$
mots creux	*avec*	*sans*	*avec*	*sans*	*avec*	*sans*	*avec*
V_antiq	86.5	80.3	76.6	75.6	86.1	87.6	90.5
WP_arch	72.9	68.2	64.2	69.2	74.5	75.2	78.5
CT_1GM	66.8	64.7	61.0	66.4	70.5	72.2	74.7
WP_1GM	74.6	70.9	65.2	64.6	73.0	73.7	76.5

TABLE 3 – Pourcentage d'appariements corrects, selon la métrique de similarité utilisée

5 Conclusions

Nous avons proposé une méthode de génération semi-automatique de questions à partir d'un corpus analysé en cadres sémantiques avec l'objectif de produire un corpus pour la compréhension automatique de texte pour la langue française. Une première validation est proposée consistant à vérifier les performances d'un appariement par similarité textuelle entre les questions générées automatiquement et des questions naturelles produites par des annotateurs. Les bonnes performances obtenues ouvrent de nombreuses perspectives. En particulier, la représentation sémantique explicite des questions obtenues, inhérente à la méthodologie de construction, permettra d'envisager des méthodes d'appariement dépassant le cadre habituel de la représentation en sacs de mots et de mesurer l'apport de telles représentations. Enfin, l'apprentissage de modèles de compréhension de texte à partir des questions générées automatiquement afin d'en vérifier la validité sur des questions naturelles sera également une piste de recherche intéressante. L'ensemble des données collectées et générées sera mis à disposition de la communauté scientifique.

Références

BAKER C. F., FILLMORE C. J. & LOWE J. B. (1998). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, p. 86–90 : Association for Computational Linguistics.

BÉCHET F., DAMNATI G., HEINECKE J., MARZINOTTO G. & NASR A. (2017). CALOR-Frame : un corpus de textes encyclopédiques annoté en cadres sémantiques. In *ACor4French – Les corpus annotés du français - Atelier TALN*, Orléans, France.

BOYER K. E. & PIWEK P. (2010). *Proceedings of QG2010 : The Third Workshop on Question Generation*. questiongeneration. org.

CHARLET D. & DAMNATI G. (2017). Simbow at semeval-2017 task 3 : Soft-cosine semantic similarity between questions for community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, p. 315–319.

CLARK P., COWHEY I., ETZIONI O., KHOT T., SABHARWAL A., SCHOENICK C. & TAFJORD O. (2018). Think you have solved question answering ? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv :1803.05457*.

DONG X., HONG Y., CHEN X., LI W., ZHANG M. & ZHU Q. (2018). Neural question generation with semantics of question type. In *CCF International Conference on Natural Language Processing and Chinese Computing*, p. 213–223 : Springer.

DUNN M., SAGUN L., HIGGINS M., GUNEY V. U., CIRIK V. & CHO K. (2017). Searchqa : A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv :1704.05179*.

FLOR M. & RIORDAN B. (2018). A semantic role-based approach to open-domain automatic question generation. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, p. 254–263.

HERMANN K. M., KOČISKÝ T., GREFENSTETTE E., ESPEHOLT L., KAY W., SULEYMAN M. & BLUNSOM P. (2015). Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, p. 1693–1701, Cambridge, MA, USA : MIT Press.

KOČISKÝ T., SCHWARZ J., BLUNSOM P., DYER C., HERMANN K. M., MELIS G. & GREFENSTETTE E. (2018). The narrative qa reading comprehension challenge. *Transactions of the Association of Computational Linguistics*, **6**, 317–328.

MAZIDI K. & NIELSEN R. D. (2014). Linguistic considerations in automatic question generation. In *ACL*.

NGUYEN T., ROSENBERG M., SONG X., GAO J., TIWARY S., MAJUMDER R. & DENG L. (2016). Ms marco : A human generated machine reading comprehension dataset. *arXiv preprint arXiv :1611.09268*.

PILLAI L. R., VEENA G. & GUPTA D. (2018). A combined approach using semantic role labelling and word sense disambiguation for question generation and answer extraction. In *2018 Second International Conference on Advances in Electronics, Computers and Communications (ICAECC)*, p. 1–6 : IEEE.

RAJPURKAR P., ZHANG J., LOPYREV K. & 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*, p. 2383–2392 : Association for Computational Linguistics.

SEO M., KEMBHAVI A., FARHADI A. & HAJISHIRZI H. (2016). Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv :1611.01603*.

TRISCHLER A., WANG T., YUAN X., HARRIS J., SORDONI A., BACHMAN P. & SULEMAN K. (2016). Newsqa : A machine comprehension dataset. *arXiv preprint arXiv :1611.09830*.

WANG S. & JIANG J. (2016). Machine comprehension using match-lstm and answer pointer. *arXiv preprint arXiv :1608.07905*.

WANG T., YUAN X. & TRISCHLER A. (2017). A joint model for question answering and question generation. *CoRR*, **abs/1706.01450**.

YUAN X., WANG T., GULCEHRE C., SORDONI A., BACHMAN P., ZHANG S., SUBRAMANIAN S. & TRISCHLER A. (2017). Machine comprehension by text-to-text neural question generation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, p. 15–25 : Association for Computational Linguistics.

ZHANG S., LIU X., LIU J., GAO J., DUH K. & VAN DURME B. (2018). Record : Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv :1810.12885*.