CALOR-QUEST : generating a training corpus for Machine Reading Comprehension models from shallow semantic annotations

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

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 and Chinese 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.

1 Introduction

Machine Reading Comprehension (MRC) is a *Natural Language Understanding* task consisting in retrieving text segments from a document thanks to a set of questions, each segment being an answer to a particular question. This task received a lot of attention in the past few years thanks to the availability of very large corpora of triplets *(document, question, answer)* such as **SQuAD** (Rajpurkar et al., 2016) or **MS MARCO** (Nguyen et al., 2016), each containing more than 100k triplets. In these corpora each question has been manually produced, either through crowd-sourcing or by collecting query logs from a search engine.

These large corpora opened the door to the development of supervised machine learning approaches for MRC, mostly based on Deep Neural Network (Wang and Jiang, 2016; Seo et al., 2016), improving greatly the state-of-the-art over previous methods based on linguistic analysis or similarity metrics between questions and segments (Hermann et al., 2015). Recently the use of contextual word embeddings such as **BERT** (Devlin et al., 2018) or **XLNet** (Yang et al., 2019) lead to obtain another great increase in performance, reaching human-level performance according to some benchmarks [1].

These large corpora are only available in English, and more recently Chinese (He et al., 2018) but for other languages, such as French, there is no comparable resources and the effort required to collect such a large amount of data is very important, limiting the use of these methods to other languages or other application frameworks.

To address this problem, several studies have proposed to *generate* automatically questions and answers directly from a text document such as Wikipedia pages (Du and Cardie, 2018) in order to build a training corpus for MRC models. One of the issues of such methods is the semantic errors that can occur between questions and answers due to the automatic generation process.

In order to try to overcome this problem, the method proposed in this paper makes use of a *FrameNet* semantic analysis of the documents in order to automatically generate questions. Semantic annotations are used to control the question generation process and the answer span identification.

We present in this study the CALOR-QUEST corpus which contains almost $100K$ triplets (text, question, answer) automatically obtained on French encyclopedic documents (Wikipedia, Vikidia, ClioTexte) with our semantically controlled question generation method. We report results on an MRC task similar to SQuAD obtained with the BERT SQuAD model (Devlin et al., 2018) fine-tuned on CALOR-QUEST and evaluated on a ma-

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

Proceedings of the Second Workshop on Machine Reading for Question Answering, pages 19–26
Hong Kong, China, November 4, 2019. ©2019 Association for Computational Linguistics

nually collected corpus in French .

2 Related work

In addition to SQuAD and MS-MARCO already mentioned in the introduction, several corpora in English have been proposed for MRC tasks as presented in (Nguyen et al., 2016), such as **NewsQA** (Trischler et al., 2016), **SearchQA** (Dunn et al., 2017) including questions from the Jeopardy game paired to text segments collected through search queries, **NarrativeQA** (Kočiský et al., 2018) built from films and books abstract.

Developing such resources for a new language requires a lot of effort, as presented in (He et al., 2018) for Chinese. In this context, methods that can help reducing this cost have attracted a lot of attention and can be grouped into two categories : methods based on an automatic *translation* process between MRC resources in English and the target language ; methods based on an automatic *question generation* and *answer spans identification* process directly from documents in the target language.

In the first category, 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) or to build a multilingual model with a GAN-based approach in which English and target language features can be joined (Lee and Lee, 2019). All these methods imply that the models or the resources created on the target language are on the same domains than the source language ones.

The second category of methods is more generic as it can be applied to any language or any domain, however it is more challenging since there is no human supervision used in the pairing of questions and answer spans.

Question generation from text has been the subject of many studies outside the scope of MRC, for example through evaluation programs such as (Boyer and Piwek, 2010). Traditionally two kinds of methods have been explored, whether through patterns built from the syntactic parsing of a sentence or from semantic analysis (Yao et al., 2012). Recent advances in these two fields have led to further advances in question generation (Mazidi and Nielsen, 2014). Recently, for example, (Pillai et al., 2018) and (Flor and Riordan, 2018) have proposed to generate factual questions from an analysis in PropBank semantic roles.

However these works often take place in an application context very different from MRC, namely the production of questions for language learning or quiz generation for education. In such contexts, the readability and grammaticality of the questions obtained is paramount and questions are usually evaluated by subjective tests or metrics like *BLEU* or *Meteor*.

Beyond knowledge-based pattern-based approaches, recent work consider question generation as a supervised machine learning task where questions or question patterns are generated by an end-to-end neural network directly from text (Dong et al., 2018; Yuan et al., 2017; Duan et al., 2017) conditioned by answer spans, even considering jointly question generation and answer span identification (Wang et al., 2017). In (Du and Cardie, 2018), the SQuAD corpus is used to train a question generation model that first extract candidate answers from Wikipedia documents, then generate answer-specific questions. This model takes co-reference into account, allowing to produce questions spanning over several sentences, a very important feature considering that nearly 30% of human-generated questions in SQuAD rely on information beyond a single sentence (Du and Cardie, 2017).

The question corpus generated by such approaches can then be used to train a MRC model, however there are two drawbacks with these methods : firstly the need for a large question/answer corpus in order to train question generation models, although such resource is not available for every language, especially for French which is the focus of this study ; secondly the fact that semantic errors can occur in the question/answer-span generation process, leading to introduce noise in the training corpus. One way to control this noise is to use an explicit semantic representation in order to relate questions and answers. This was done in (Serban et al., 2016) by using the *Freebase* (Bollacker et al., 2008) knowledge base combined to a question dataset (*SimpleQuestion* dataset (Bordes et al., 2015)) in order to generate a very large corpus of questions on the *Freebase* entities and relations.

The approach followed in this study is also based on an explicit semantic representation in order to generate pairs of question/answer-span. The main difference is that since we don't have a large corpus of question/answer pairs to train a question generation model, we will rely on simple patterns based on the semantic annotations of our target corpus. The main originality of this work is to use a large encyclopedic corpus in French annotated with a FrameNet semantic model, the **CALOR-FRAME** corpus (Marzinotto et al., 2018), in order to automatically produce a large amount of semantically-valid pairs of questions and answer-spans, the **CALOR-QUEST** corpus.

Using FrameNet annotations for generating an MRC training corpus has a major drawback : the human effort needed to build such resources is arguably bigger than building directly a question/answer corpus such as SQuAD. However we believe this method has several advantages :

- firstly corpora with frame-based annotations are available for many languages, even if often of limited sizes ;
- secondly frame-based annotation is not linked to a single task such as MRC, therefore data developed for other application frameworks can be reuse ;
- lastly the availability of *explicit* semantic annotations on which an end-to-end MRC model is trained and evaluated can give us insights about what is being learned by these models and on their generalization capabilities, as our first experiments will show in section 4.2.

3 Using shallow semantic annotations to obtain a question/answer corpus

The **CALOR-FRAME** corpus is made of 4 sub-corpora stemming from 3 encyclopedic sources : Wikipedia (WP), Vikidia (V) and ClioTexte (CT). Three themes are covered : World War I (WWI), archaeology (arch) and antiquity (antiq). This variety spans different genres ranging from historical documents for ClioTexte (speeches, declarations) to article for children in Vikidia. The corpus was hand-annotated with Semantic Frames, following the *Berkeley FrameNet* (Baker et al., 1998) annotation guidelines. Semantic Frames describe prototypical situations, such as *decide, lose, attack, win*. Every Frame has a *Lexical Unit (LU)*, which is a word or an expression that triggers the Frame and *Frame Elements (FE)* which are the participants to the situation denoted by the Frame. Every FE has a *label*, such as *Agent, Patient, Time, . . .* that denotes the relation that links the FE and the Frame.

In the **CALOR-FRAME** corpus, 54 different Frames were used, that can be triggered by 145 lemmas (70 nouns and 75 verbs), as described in (Béchet et al., 2017). The annotation process of a sentence consists in first identifying the potential triggers, then the Frame triggered, and their FEs. A sequence of words can correspond to several FEs for sentences with several Frames occurrences. An example is given in Figure 1 for a sentence with two Frame occurrences : an occurrence of the Frame `Losing` triggered by the word `lost` and an occurrence of the Frame `Attack` triggered by the noun `attacks`[2].

When a FE is a pronoun (*e.g. they*) or a sub-specified noun phrase (*e.g. the troops*) the co-reference to the explicit mention (*e.g. German troops*) is annotated, therefore a Frame can span other several sentences in a document.

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, E is one Frame Element of f and C (for Context) is the set of the other Frame Elements. Given a triplet (F, E, C), questions can be produced for which the answer is E.

In the case of the Losing Frame of Figure 1, which has three Frame Elements, three triplets (F, E, C) can be produced :

```
(Losing, Owner, {Time, Possession})
(Losing, Time, {Owner, Possession})
(Losing, Possession, {Owner, Time})
```

When a frame element is a co-reference to an explicit mention, it is the mention which is used, therefore a question can spread over several sentences in a document.

A triplet (F, E, C) yields a set of questions noted *Questions*(F, E, C). The first triplet in the above list, for example, can produce the question *Who lost 80% of its number on 8 and 9 October.* or *Which troops were wiped out during the attacks of early October ?* Both questions have as an answer the same text segment : *the German troops*, but present very different forms : the first one is close to the original sentence and could be produced by a simple re-organization of the latter, without any lexical change while the second asks for a complete rewriting. Both types of questions have been produced using the Frame annotation of the

[2] The original example is in French, it has been translated for readability reasons.

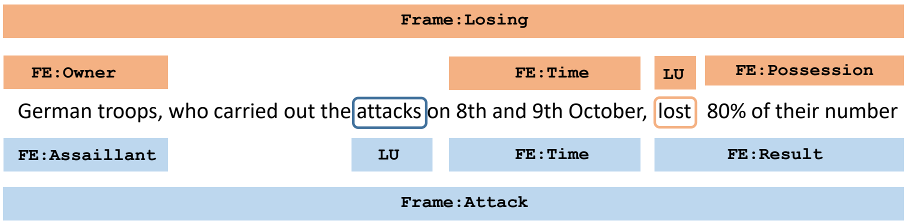
Frame:Losing
FE:Owner
FE:Time
LU
FE:Possession
German troops, who carried out the attacks on 8th and 9th October, lost 80% of their number
FE:Assailant
LU
FE:Time
FE:Result
Frame:Attack

Figure 1: A sentence annotated with Frames defined in the Berkeley Framenet project

CALOR-FRAME corpus. The first type, called *automatic questions*, noted Q_A, has been produced automatically using rules applied on (F, E, C) triplets. The second type, called *natural questions*, noted Q_N, has been produced manually using a sub-part of the corpus, in a controlled setup. Both methods are described below.

In both cases, hand-produced Frame annotations have been used, as a proof of concept of the proposed method. Question production based on automatically predicted Frames is left for future work.

3.1 Rule-based question generation

The automatic production of questions is based on rules which are sentences with variables that correspond to FE. When applied on a triplet, the variables are instantiated with the corresponding FE. Some variables are optional, they can be omitted in the question. A single rule applied to a (F, E, C) triplet can therefore produce several questions.

Two rules are represented in Figure 2. Variables are prefixed with a $ sign, followed by an FE label. Optional parts are represented between square brackets.

F = Leadership E = Time	
gen.	When lead [$Leader] [$Governed] [$Place] $Role] [$Duration] [$Activity] ?
spe.	When did $Leader lead $Governed [$Place] ?

Figure 2: Example of a generic and a specific rule for the Leadership Frame and the FE Time

Two types of rules has been used for generating questions : generic rules and specific rules.

Generic rules are produced automatically from a Frame description F, the indication of a specific FE E that corresponds to the answer of the question as well as the set of all possible verbs that can trigger F. The rules are built by selecting an interrogative pronoun that is compatible with E[3] followed by a possible trigger for F and all possible combinations of FE excluding E which is the answer to the question. In the example of the generic rule of Figure 2, the pronoun is *When*, the trigger is *lead*, followed by all FE except the FE Time, which is the answer. Every FE is optional, which allows to exclude any subset of FE from the question. Such rules can lead to awkward questions, due either to lexico-syntactic reasons or to the choice of optional FE that are kept in the question produced.

Specific rules are built manually. They share the same format as generic rules but there is a manual control on the lexical and syntactic aspects of the question as well as the FE that are considered mandatory or optional.

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.

3.2 Collecting *real* questions from semantic annotations

To obtain *real* questions for our evaluation corpus we could have used the same protocol as for SQuAD and ask annotators to produce arbitrary questions directly from the CALOR-FRAME corpus. However, as discussed in section 2, one of the goals of this study is to provide insights about what is being learned by MRC end-to-end models by controlling semantic of both training and evaluation data. Therefore we decided to produce natural questions with annotators to whom (F, E, C) triplets were shown. The original sentence was not presented in order to leave more freedom for the annotator in her or his lexical ans syntactic choices. Besides, the annotator can select any elements of the context to include in the question. The

[3] The list of compatible pairs of an interrogative pronoun and a FE has been built manually.

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.

The following example shows in the upper part the information that were given to the annotator and in the lower part, some questions produced.

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 ?*

The natural questions produced with this protocol concerned only a sub-part of the CALOR-QUEST corpus but this sub-part has been selected in order to represent all the Frames used to annotate the corpus.

3.3 Collected corpus

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 1.

collection	#docs	*#natural* questions	*#generated* questions
V_antiq	61	274	4672
WP_arch	96	302	36259
CT_1GM	16	241	7502
WP_1GM	123	319	50971
total	**296**	**1136**	**99404**

Table 1: Description of CALOR-QUEST corpus

4 Evaluation

The main objective of our work is to create in a semi-automatic fashion a training corpus for reading comprehension model. Thus, for a given document annotated with frames, we want to generate automatically as many questions as possible, semantically valid, for which we have, by construction, the right answer span in the document. To validate this approach we perform an extrinsic evaluation of this corpus by using it for training a state-of-the-art Machine Reading Comprehension system publicly available, and by evaluating its performance on the set of natural questions collected.

4.1 Experiments with BERT-SQUAD

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)[4], with default hyperparameters. To be in 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 evalution script (`https://github.com/allenai/bi-att-flow/blob/master/squad/evaluate-v1.1.py`), customized for French (removing french articles in the normalization process, instead of english articles). In this evaluation set-up, "*Exact Match*" represents the percentage of questions whose predicted answer matches exactly the ground-truth answer, and "*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.

The training is done on a randomly selected sample set of 14K generated questions (due to our computational storage limitation). In these experiments we first select automatic questions generated thanks to specific rules, then add questions produced by generic rules. The evaluation is done on the natural questions set. For the *SQUAD1.1* condition, all the questions are answerable in the given paragraphs. For the *SQUAD2.0* condition, the system is also able to detect if a question is answerable or not, in a given paragraph. For this set up, we build a specific test set, with 2/3 made of answerable questions for a given paragraph, and 1/3 made of answerable questions of another paragraph of the same document, thus assumed to be

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

unanswerable for the given paragraph (but dealing with the same topic).

Results are presented in table 2. As we can see the model find the correct answer with the correct span for about 60% of the natural questions of our test corpus, although it has been trained only on generated questions. F1 measure is satisfying for SQUAD1.1i (76.7), although it drops to 64.2 when introducing unanswerable questions in SQUAD2.0. However performance of unanswerable question detection are excellent (98.0). This validate our approach although there is a large margin of improvement considering the performance of current models on the English SQUAD corpus.

version	exact	F1	F1-HasAns	F1-NoAns
V1	59.4	76.7	76.7	-
V2	62.7	73.5	64.2	98.0

Table 2: Results obtained with BERT-SQUAD on CALOR-QUEST with two conditions : V1 correspond to SQUAD1.1 where all questions refer to an answer in the documents; V2 correspond to SQUAD2.0 with unanswerable questions

4.2 Contrastive experiments

4.2.1 Generalization beyond the initial semantic model

In the following experiment, we evaluate how this framework generalizes to new semantic frames, and is able to answer questions related to semantic frames which were absent from the training set. To this purpose, we select 10 semantic frames for which we have the most numerous natural questions in the test set, and we discard from the training set the questions generated from these 10 semantic frames. In table 3, performances are reported for each subset of natural questions, including or not generated questions from the same semantic frame in the training. It can be seen that for most of these frames the decrease of performance observed when excluding them from the training set is not important, therefore we can conclude that our method allows to train models that can generalize beyond the set of semantic frames that was used to generate the training corpus.

However table 3 also shows that for 3 of these frames (*Departing, Appointing, Shoot-projectiles*) there is a loss of more than 10% F1 when using the reduced training corpus, indicating that this generalization capabilities can be limited for some specific actions.

Frame	#quest	F1 (all)	F1 (w/o)
Death	49	89.2	78.78
Creating	38	81.0	85.0
Existence	38	83.0	79.2
Giving	65	80.9	73.8
Coming-up-with	27	86.4	82.9
Departing	64	86.3	71.8
Appointing	52	76.9	64.5
Buildings	62	77.5	73.6
Colonization	37	69.9	70.0
Shoot-projectiles	24	72.4	47.3

Table 3: F1 results on questions associated to 10 semantic frames of CALOR-QUEST with a model trained on the whole corpus (*all*) and one trained on a corpus where all the generated questions corresponding to these 10 frames have been removed (*w/o*)

4.2.2 Generalization to a new domain

We also evaluate our method on a different corpus to check how domain-depend are the models trained on **CALOR-QUEST** . (Asai et al., 2018) provides a French transcription of a subset of the development set of the original SQuAD corpus. They have extracted several paragraphs and their corresponding questions, resulting in 327 paragraph-question pairs over 48 articles. This subcorpus was manually translated into French by bilingual workers on Amazon Mechanical Turk and further corrected by bilingual experts. In this corpus of 327 questions, only 46 correspond to questions related to frames defined in CALOR, and themes are not restricted to historical knowledge as in CALOR. Thus, we have a semantic shift but also a lexical shift between the training set of CALOR and the testing set of french_squad.

We test this corpus on the model trained on **CALOR-QUEST** . For sake of comparison, we report baseline performance obtained in (Asai et al., 2018) with a back-translation approach : the French evaluation corpus is first translated to English with an automatic French-to-English translation service, then the BERT system with English model is applied to this automatic translation, finally the outputs of the system are automatically back-translated to French for evaluation.

Results are presented in table 4. Although we observe an important decrease in performance, in comparison with the results obtained on CALOR, performance is still much better than the one obtained with a back-translation baseline of the well-trained BERT-model in English.

Model	exact	F1
CALOR-QUEST	38.5	53.6
BERT-SQUAD (auto trans.) (Asai et al., 2018)	23.5	44.0

Table 4: Results obtained on the French SQuAD test corpus with a model trained on CALOR-QUEST and the original BERT-SQUAD model for English with back-translation

5 Conclusion

In this work, we have proposed a semi-automatic method to generate question/answer pairs from a corpus of documents annotated in semantic frames, with the purpose of building a large training corpus for machine reading comprehension in French. Based on simple rules applied on shallow semantic annotations, the produced questions are valid semantically, but their syntactic validity is not guaranteed. Additionally, a set of more than 1000 question/answer pairs has been collected manually, to be used as a test corpus. We validate the usefulness of the corpus of automatic questions, by training a state of the art, publicly available, machine reading comprehension system, based on fine-tuning multilingual BERT features on this corpus. We then test the resulting model on the set of real questions, and on a french translation of a subset of the SQuAD corpus, and promising results have been obtained. Further work will focus on extending this approach to semantic annotations obtained automatically. The extension to another semantic annotation scheme such as PropBank will also be studied. The **CALOR-QUEST** corpus of automatic and natural questions will be made publicly available, to foster machine reading comprehension for French language.

References

Akari Asai, Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2018. Multilingual extractive reading comprehension by runtime machine translation. *CoRR*, abs/1809.03275.

Collin F Baker, Charles J Fillmore, and John B Lowe. 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*, pages 86–90. Association for Computational Linguistics.

Frédéric Béchet, Géraldine Damnati, Johannes Heinecke, Gabriel Marzinotto, and Alexis Nasr. 2017. CALOR-Frame : un corpus de textes encyclopédiques annoté en cadres sémantiques. In *ACor4French – Les corpus annotés du français - Atelier TALN*, Orléans, France.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase : a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. AcM.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *arXiv preprint arXiv :1506.02075*.

Kristy Elizabeth Boyer and Paul Piwek. 2010. *Proceedings of QG2010 : The Third Workshop on Question Generation*. questiongeneration. org.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

Xiaozheng Dong, Yu Hong, Xin Chen, Weikang Li, Min Zhang, and Qiaoming Zhu. 2018. Neural question generation with semantics of question type. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 213–223. Springer.

Xinya Du and Claire Cardie. 2017. Identifying where to focus in reading comprehension for neural question generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2067–2073.

Xinya Du and Claire Cardie. 2018. Harvesting paragraph-level question-answer pairs from wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1907–1917.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 866–874, Copenhagen, Denmark. Association for Computational Linguistics.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa : A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv :1704.05179*.

Michael Flor and Brian Riordan. 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*, pages 254–263.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, et al. 2018. Dureader : a chinese machine reading comprehension dataset from real-world applications. *ACL 2018*, page 37.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 1693–1701, Cambridge, MA, USA. MIT Press.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gáabor Melis, and Edward Grefenstette. 2018. The narrative qa reading comprehension challenge. *Transactions of the Association of Computational Linguistics*, 6 :317–328.

Chia-Hsuan Lee and Hung-Yi Lee. 2019. Cross-lingual transfer learning for question answering. *arXiv preprint arXiv :1907.06042*.

Gabriel Marzinotto, Jeremy Auguste, Frederic Bechet, Géraldine Damnati, and Alexis Nasr. 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.

Karen Mazidi and Rodney D. Nielsen. 2014. Linguistic considerations in automatic question generation. In *ACL*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco : A human generated machine reading comprehension dataset. *arXiv preprint arXiv :1611.09268*.

Lekshmi R Pillai, G Veena, and Deepa Gupta. 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)*, pages 1–6. IEEE.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 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.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv :1611.01603*.

Iulian Vlad Serban, Alberto García-Durán, Caglar Gulcehre, Sungjin Ahn, Sarath Chandar, Aaron Courville, and Yoshua Bengio. 2016. Generating factoid questions with recurrent neural networks : The 30m factoid question-answer corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 588–598.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa : A machine comprehension dataset. *arXiv preprint arXiv :1611.09830*.

Shuohang Wang and Jing Jiang. 2016. Machine comprehension using match-lstm and answer pointer. *arXiv preprint arXiv :1608.07905*.

Tong Wang, Xingdi Yuan, and Adam Trischler. 2017. A joint model for question answering and question generation. *CoRR*, abs/1706.01450.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Xuchen Yao, Gosse Bouma, and Yi Zhang. 2012. Semantics-based question generation and implementation. *Dialogue & Discourse*, 3(2) :11–42.

Xingdi Yuan, Tong Wang, Caglar Gulcehre, Alessandro Sordoni, Philip Bachman, Saizheng Zhang, Sandeep Subramanian, and Adam Trischler. 2017. Machine comprehension by text-to-text neural question generation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 15–25. Association for Computational Linguistics.