Clozer: Adaptable Data Augmentation for Cloze-style Reading Comprehension

Holy Lovenia[*], **Bryan Wilie**[*], **Willy Chung**[*], **Min Zeng**[*],
Samuel Cahyawijaya, **Su Dan**, **Pascale Fung**
Center for Artificial Intelligence Research (CAiRE)
The Hong Kong University of Science and Technology
(hlovenia, bwilie, whcchung, min.zeng)@connect.ust.hk

Abstract

Task-adaptive pre-training (TAPT) alleviates the lack of labelled data and provides performance lift by adapting unlabelled data to downstream task. Unfortunately, existing adaptations mainly involve deterministic rules that cannot generalize well. Here, we propose Clozer, a sequence-tagging based cloze answer extraction method used in TAPT that is extendable for adaptation on any cloze-style machine reading comprehension (MRC) downstream tasks. We experiment on multiple-choice cloze-style MRC tasks, and show that Clozer performs significantly better compared to the oracle and state-of-the-art in escalating TAPT effectiveness in lifting model performance, and prove that Clozer is able to recognize the gold answers independently of any heuristics.

1 Introduction

Endowing machines with the proficiency to read, understand, and reason from unstructured text information is an ongoing aspiration in natural language processing. This aim raises a notable research focus: machine reading comprehension (MRC). Given a question, the goal of MRC is to infer the correct answer based on important cues gathered through understanding relevant context passage. MRC tasks vary in structure, depending on their question construction (e.g., cloze-style) and answer type (e.g., multiple-choice) (Zeng et al., 2020).

Various methods using large pre-trained language models (LMs) have been proposed in MRC tasks. In recent years, adaptation methods such as task adaptive pre-training (TAPT) have been widely adopted for MRC tasks (Xie et al., 2021; Wang et al., 2021; Glass et al., 2020). TAPT uses in-domain unlabelled data of the downstream task to generate a synthetic pre-training dataset adapted to the downstream task through certain data augmentation methods, depending on the downstream task in use. For multiple-choice cloze-style MRC, data augmentation often involves two steps: 1) answer extraction or selection and 2) pseudo-answer generation (Figure 1). Both steps have been adopted in several studies with varying implementations (Welbl et al., 2017; Onishi et al., 2016; Yang et al., 2020). One notable work presents TAMAMC (Gururangan et al., 2020), which achieves state-of-the-art performance by adopting the TAPT framework. However, this method relies heavily on the downstream task's heuristics in the answer selection step, which hinders its applicability to other multiple-choice cloze-style MRC tasks.

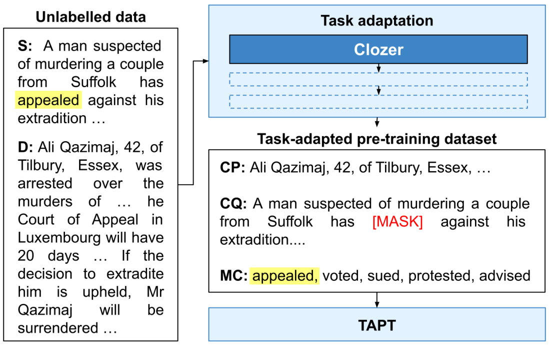
Unlabelled data
S: A man suspected of murdering a couple from Suffolk has appealed against his extradition …
D: Ali Qazimaj, 42, of Tilbury, Essex, was arrested over the murders of … he Court of Appeal in Luxembourg will have 20 days … If the decision to extradite him is upheld, Mr Qazimaj will be surrendered …
Task adaptation
Clozer
Task-adapted pre-training dataset
CP: Ali Qazimaj, 42, of Tilbury, Essex, …
CQ: A man suspected of murdering a couple from Suffolk has [MASK] against his extradition....
MC: appealed, voted, sued, protested, advised
TAPT

Figure 1: Clozer extracts an answer for TAPT

In this paper, we take a step towards generalized synthetic pre-training dataset construction, to use TAPT to solve multiple-choice cloze-style MRC. We propose Clozer, a cloze answer extraction based on sequence tagging developed independently of pre-defined rules to improve the generalizability of the TAPT method for the cloze-style MRC tasks. Clozer learns the intrinsic pattern of the downstream task dataset and acts as an answer extractor for the unlabelled data (Figure 1). To adapt to the downstream task, the extractions are grouped with several other options to form a triplet of {*context passage, cloze question, multiple-choice options*}, following the standard multiple-choice cloze-style MRC task format, as a synthetic sample for the second pre-training phase. We conduct our experiments on two downstream tasks. Our experimental results show that employing Clozer in TAPT provides a substantial performance boost, while being generally applicable for both multiple-choice MRC tasks we experiment on.

[*] The authors contributed equally to this work.

Proceedings of the 7th Workshop on Representation Learning for NLP, pages 60 - 66
May 26, 2022 ©2022 Association for Computational Linguistics

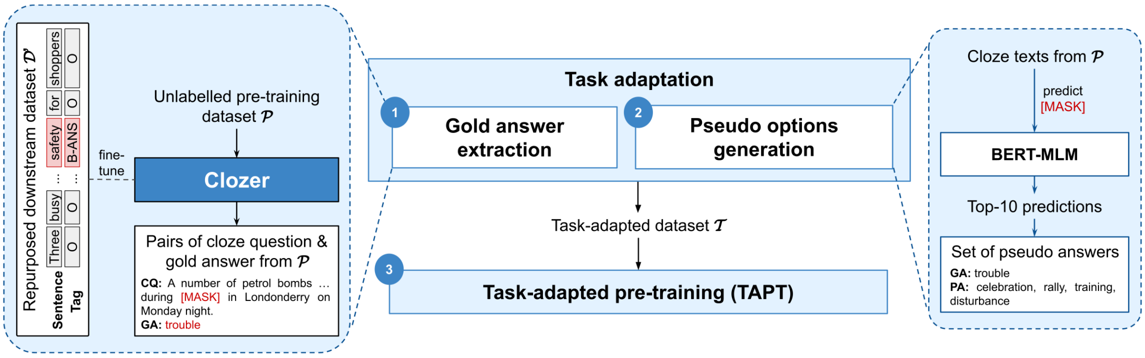
Repurposed downstream dataset $\mathcal{D}'$
Sentence
Tag
Three
busy
…
safety
for
shoppers
O
O
…
B-ANS
O
O
fine-
tune
Unlabelled pre-training
dataset $\mathcal{P}$
Clozer
Pairs of cloze question &
gold answer from $\mathcal{P}$
CQ: A number of petrol bombs … during [MASK] in Londonderry on Monday night.
GA: trouble
Task adaptation
1
Gold answer
extraction
2
Pseudo options
generation
Task-adapted dataset $\mathcal{T}$
3
Task-adapted pre-training (TAPT)
Cloze texts from $\mathcal{P}$
predict
[MASK]
BERT-MLM
Top-10 predictions
Set of pseudo answers
GA: trouble
PA: celebration, rally, training, disturbance

Figure 2: Method pipeline for Clozer-based TAPT

Our contributions are as follows: 1) to the best of our knowledge, we are the first to introduce an automatic generalizable cloze answer extraction method to support a generalized TAPT method for multiple-choice cloze-style MRC tasks; 2) we show that Clozer significantly outperforms all other baselines on two multiple-choice cloze-style MRC tasks without relying on any task-specific heuristics; and 3) we present further analysis to explain the effectiveness and efficiency of our Clozer and provide insight on how to improve its generalizability.

2 Related Work

Task-adaptive pre-training Howard and Ruder (2018) propose Universal Language Model Fine-tuning (ULMFiT), which pre-trains an LM on a large general-domain corpus and fine-tunes it on the target task. Second-phase pre-training has been used to improve the performance of an LM for certain downstream tasks such as text classification (Sun et al., 2019). Studies on TAPT (Gururangan et al., 2020; Pruksachatkun et al., 2020) prove that the performance boost it obtains can be on par with domain-adaptive pre-training, with the benefit of using a much smaller but relevant corpus. TAPT has proved effective in many downstream tasks such as abstractive summarization (Yu et al., 2021) and dialogue systems (Zhang et al., 2021a).

Answer extraction Tan et al. (2018) develop an extraction-then-synthesis framework to synthesize answers from extraction results. Specifically, the answer extraction model is first employed to predict the most important sub-spans from the passage, then the answer synthesis model takes the sub-spans as additional features along with the question and passage to further elaborate the final answers. Xiong et al. (2016) introduce the Dynamic Coattention Network (DCN) for a question-answering task, which learns the co-dependent representations of the question and the passages. Seo et al. (2016) introduce the Bi-Directional Attention Flow (BIDAF) network to match the question and passages. It uses the BIDAF mechanism to get a query-aware context representation without early summarization.

Sequence tagging Sequence tagging is utilized to assign a label for each token (i.e., word) in a sequence. While it's commonly applied for tasks like named entity recognition (NER), part-of-speech (POS) tagging, and text chunking, Yao et al. (2013); Wilie et al. (2020) prove that it is feasible to use this approach to construct cloze questions by extracting an answer span from a complete sentence. Yao et al. (2013) cast answer extraction as an answer sequence-tagging task, utilizing a linear-chain conditional random field (CRF) with tree edit distance (TED) and traditional contextual features.

3 Methodology

Our method follows the pipeline described in Figure 2. We follow TAPT's objective, in which a model learns on a small task-relevant set of data instead of doing another round of masked language modeling (MLM) for pre-training. Utilizing Clozer, we adapt a large unlabelled pre-training dataset based on the downstream task, which could be any multiple-choice cloze-style MRC task.

We define the pre-training dataset $\mathcal{P} = \{(d_i^{\mathcal{P}}, s_i^{\mathcal{P}})\}_{i=1}^{n}$ with $d_i^{\mathcal{P}}$ as a document and $s_i^{\mathcal{P}}$ as a summary or a single sentence related to the passage $d_i^{\mathcal{P}}$. $\mathcal{P}$ could be any unlabelled data of document and sentence pairs, e.g., headline-content of news, title-body of articles, and synopsis-narration of stories. Through the task adaptation, we reconstruct $\mathcal{P}$ into a synthetic cloze-style MRC task, where the resulting task-adapted pre-training dataset is represented by $\mathcal{T} = \{(c_i^{\mathcal{T}}, q_i^{\mathcal{T}}, o_i^{\mathcal{T}}, l_i^{\mathcal{T}})\}_{i=1}^{m}$. It follows the structure of the downstream task dataset $\mathcal{D} = \{(c_i^{\mathcal{D}}, q_i^{\mathcal{D}}, o_i^{\mathcal{D}}, l_i^{\mathcal{D}})\}_{i=1}^{m}$, where $c_i^{\mathcal{D}}$ is a context passage, $q_i^{\mathcal{D}}$ is a cloze question, $o_i^{\mathcal{D}} \in o_1, \ldots, o_k$ is a set of multiple-choice options, and $l_i^{\mathcal{D}}$ is the gold answer's index as the correct label.

We split the task adaptation into 1) gold answer extraction and 2) pseudo options generation, which are explained in §3.1 and §3.2 respectively. Afterwards, we employ TAPT using the task-adapted dataset $\mathcal{T}$, the details of which are provided in §3.3.

3.1 Gold answer extraction

Gold answer extraction (GAE) represents the pre-training dataset's summary as a cloze question by taking out a gold answer, which depends on the downstream task's notion of what is a correct answer. We tackle this problem by utilizing Clozer to learn from the downstream task and identify the appropriate gold answers by sequence tagging. First, we repurpose the cloze questions and gold answers in the downstream task as a token classification dataset. We use the tag `B-ANS` for the gold answer and the tag `O` for other words in the cloze question.

Afterwards, we fine-tune Clozer on this repurposed dataset so it can learn and approximate the downstream task's pattern of determining the gold answers. It is worth mentioning that, due to its independence from any heuristic rules, our Clozer method is not constrained to a single specific interpretation of gold answers. It can be adapted to extract any type of cloze answers (e.g., abstract meaning) depending on the downstream task dataset. We next use Clozer to predict the pre-training dataset's summaries and extract the gold answers. We replace the gold answers in the summaries with the `[MASK]` token to form cloze questions and pass the questions on to the next step. We drop candidates with zero or more than one gold answer.

3.2 Pseudo options generation

Pseudo answer generation (POG) employs a pre-trained masked LM to predict the `[MASK]` token. For each cloze question, we obtain the model's top predictions and filter out the ones that are incomplete or too similar to the gold answer. We randomly pick k predictions as pseudo options. We discard data samples with fewer than k remaining predictions. After this step, each pre-training dataset sample consists of a context paragraph, a cloze question, a gold answer, and four pseudo options. Following the downstream task dataset structure, we recast the gold answer and pseudo options as $\{o_1, o_2, \ldots, o_k\}$ in random order. The gold answer's option index becomes the label. In cases beyond the scope of this work where multiple-choice is not required by the cloze task, POG is skipped.

3.3 Task-adaptive pre-training

We feed the task-adapted dataset to a pre-trained multiple-choice classification model for TAPT. The final step is to fine-tune the model on the downstream task and evaluate it. To see how Clozer performs against other available methods, we present the results of three baselines, where we employ a directly fine-tuned model, TA-MAMC, and an oracle in place of Clozer in the GAE step. The baselines will be further explained in §4.

4 Experiment

Dataset As explained in §3, the methodology requires the usage of a pre-training dataset and a downstream task. In the experiment, we apply Clozer for the TAPT method on two downstream tasks separately. Both are multiple-choice cloze-style MRC tasks and are obtained from the subtask 1 and subtask 2 of ReCAM (Zheng et al., 2021). Given a context passage and multiple choice options, the appropriate gold answer must be derived to complete a cloze question. The first task defines its gold answers as imperceptible concepts, while the second defines them as hypernyms. For the pre-training dataset $\mathcal{P}$, we use XSUM (Narayan et al., 2018), an abstractive news summarization dataset.

Baseline To see how Clozer-based TAPT performs against other methods, we employ three baselines for the experiment: **1) direct fine-tuning**, where a pre-trained multiple-choice model applies no TAPT and is immediately fine-tuned on the downstream task; **2) TA-MAMC**, which selects gold answers by emulating the POS-tag distribution of the downstream task's training data; and **3) oracle**, whose answer selection is built upon heuristic rules specific to each downstream task.

The oracle utilizes a psycholinguistic database of abstract words (Coltheart, 1981) to select the *imperceptible* concepts as the gold answers in the first task. For the second task, it uses a hypernym hierarchy from WordNet (Changizi, 2008) to determine the gold answers. Both heuristics are chosen because they are used to select the original gold (i.e., correct) answers in the ReCAM dataset creation.

Approach	ReCAM 1		ReCAM 2	
	Acc	F1	Acc	F1
Direct FT	64.16%	64.15%	64.75%	64.65%
TA-MAMC†	64.99%	64.99%	67.69%	67.68%
Oracle	65.83%	65.80%	68.60%	68.50%
Clozer	**65.95%**	**65.96%**	**73.56%**	**73.45%**

Table 1: Performance comparison on the test sets of the downstream tasks. **Bold** marks the best results. †We reproduce this approach based on Zhang et al. (2021b).

Training and evaluation In the GAE, our Clozer is implemented using a pre-trained ELECTRA-base (Clark et al., 2020), while for the POG and TAPT, we initialize the model using a pre-trained BERT-base model (Devlin et al., 2019). Since only the training set and the development set of both downstream tasks are labelled, we split the original training set with a ratio of 80:20 to form a training set and a validation set. We use the development set as a test set. Accuracy and F1-score are used to assess the methods' performance on the test set.

5 Results and Analysis

5.1 Overall results

We present our experimental results in Table 1. Without additional TAPT, the direct fine-tuning method yields the lowest results. In comparison, TA-MAMC, which relies on POS-tag distribution, performs slightly better, and the oracle, which exploits the downstream tasks' heuristic rules, achieves the best scores among our baselines. Our proposed Clozer method, however, surpasses all baselines in both downstream tasks, by around 2% for task 1 and 9% for task 2. While Clozer provides substantial improvements, there is a considerable discrepancy between both performances due to the way the tasks are defined. We further discuss Clozer's performance discrepancy in §5.3.

5.2 Quality of answer extraction methods

As shown in Table 2, the oracle, which derives its understanding of the answers from the semantics provided by the heuristic rules, has the fewest data after the GAE step (94k out of 200k), because the heuristic rules it is built upon are deterministic and leave no room for randomness. TA-MAMC's POS-tag distribution approach provides some knowledge of the target's syntax but represents no semantic ties to ReCAM's answers (i.e., imperceptible concepts and hypernyms).

Task adapter	ReCAM 1		ReCAM 2	
	Post-GAE	Post-POG	Post-GAE	Post-POG
TA-MAMC	155017	47699	155858	48358
Oracle	94954	29073	75920	23520
Clozer	120073	35073	181476	53368

Table 2: Number of data samples left after the **GAE** and **POG** for different task-adapter methods.

However, TA-MAMC has the benefit of excluding fewer examples than the oracle. Our Clozer finds a middle ground by being more generalizable compared to both baselines, while producing a better answer extraction quality (Table 1). Clozer shows superior results with only 5k more data samples in task 2 and with 12k fewer data samples in task 1. This shows that, while the amount of data contributes to the performance lift, the quality of the extracted answers in the synthetic task-adapted dataset is indispensable.

5.3 Clozer's performances on different downstream tasks

While TAPT lifts the model performance by 2% for task 1 and 9% for task 2, the difference between the tasks is glaring. We argue that this is largely due to the amount of synthetic data left after applying the task adaptation, as shown in Table 2, with 35k samples left in task 1 and 53k samples in task 2. This shows that the definition of abstractness chosen by ReCAM for gold answers in task 1 is more complex than the definition used by task 2, which causes the answers in task 1 to be harder to grasp by all of the approaches, including our Clozer.

This is coherent as ReCAM defines *imperceptible* concepts in task 1 using a model-based approach, which in turn introduces an innate bias to the definition. This causes identifying answers in task 1 to be conceptually more complex than in task 2, where the answers are simply nouns and verbs derived from a hypernym hierarchy. This is also in line with Zheng et al. (2021), who show that the cross-task performance drops significantly more for models trained on task 2 trying to make predictions on task 1, rather than the opposite. Examples of this complexity difference are in Appendix A.

6 Conclusion

We have proposed Clozer, an automatic generalizable cloze answer extraction method, to help in synthetic TAPT dataset construction in multiple-choice cloze-style MRC tasks. Performing TAPT with gold answers extracted by our ELECTRA-based Clozer produces stronger models than the baselines in terms of effectiveness (i.e., performance) and efficiency (i.e., the amount of data used in TAPT). Moreover, we also show that the quality of Clozer's extracted answers is higher, despite its independence from the downstream task's heuristics

Acknowledgement

This work has been supported by the China NSFC Project (No. NSFC21EG14), School of Engineering PhD Fellowship Award, the Hong Kong University of Science and Technology and PF20-43679 Hong Kong PhD Fellowship Scheme, Research Grant Council, Hong Kong.

References

Mark A. Changizi. 2008. Economically organized hierarchies in wordnet and the oxford english dictionary. *Cognitive Systems Research*, 9(3):214–228.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Michael Glass, Alfio Gliozzo, Rishav Chakravarti, Anthony Ferritto, Lin Pan, G P Shrivatsa Bhargav, Dinesh Garg, and Avi Sil. 2020. Span selection pre-training for question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2773–2782, Online. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Takashi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2016. Who did what: A large-scale person-centered cloze dataset. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2230–2235.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel Bowman. 2020. Intermediate-task transfer learning with pre-trained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247.

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

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? *CoRR*, abs/1905.05583.

Chuanqi Tan, Furu Wei, Nan Yang, Bowen Du, Weifeng Lv, and Ming Zhou. 2018. S-net: From answer extraction to answer synthesis for machine reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Ye Wang, Yanmeng Wang, Haijun Zhu, Bo Zeng, Zhenghong Hao, Shaojun Wang, and Jing Xiao. 2021. PINGAN omini-sinitic at SemEval-2021 task 4:reading comprehension of abstract meaning. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 820–826, Online. Association for Computational Linguistics.

Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857, Suzhou, China. Association for Computational Linguistics.

Xin Xie, Xiangnan Chen, Xiang Chen, Yong Wang, Ningyu Zhang, Shumin Deng, and Huajun Chen. 2021. ZJUKLAB at SemEval-2021 task 4: Negative augmentation with language model for reading comprehension of abstract meaning. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 810–819, Online. Association for Computational Linguistics.

Caiming Xiong, Victor Zhong, and Richard Socher. 2016. Dynamic coattention networks for question answering. *arXiv preprint arXiv:1611.01604*.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. Generative data augmentation for commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1008–1025.

Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. 2013. Answer extraction as sequence tagging with tree edit distance. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 858–867.

Tiezheng Yu, Zihan Liu, and Pascale Fung. 2021. Adaptsum: Towards low-resource domain adaptation for abstractive summarization. *arXiv preprint arXiv:2103.11332*.

Changchang Zeng, Shaobo Li, Qin Li, Jie Hu, and Jianjun Hu. 2020. A survey on machine reading comprehension—tasks, evaluation metrics and benchmark datasets. *Applied Sciences*, 10(21):7640.

Boliang Zhang, Ying Lyu, Ning Ding, Tianhao Shen, Zhaoyang Jia, Kun Han, and Kevin Knight. 2021a. A hybrid task-oriented dialog system with domain and task adaptive pretraining. *arXiv preprint arXiv:2102.04506*.

Jing Zhang, Yimeng Zhuang, and Yinpei Su. 2021b. TA-MAMC at SemEval-2021 task 4: Task-adaptive pretraining and multi-head attention for abstract meaning reading comprehension. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 51–58, Online. Association for Computational Linguistics.

Boyuan Zheng, Xiaoyu Yang, Yu-Ping Ruan, Zhenhua Ling, Quan Liu, Si Wei, and Xiaodan Zhu. 2021. SemEval-2021 task 4: Reading comprehension of abstract meaning. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 37–50, Online. Association for Computational Linguistics.

A Examples of Gold Selection with Clozer

David Beckham has expressed his pride at helping London win their 2012 olympics bid despite not being picked in great britains football squad.
A 22 year old man arrested on suspicion of murder following the death of Lewis Siddall has been released on bail.
A cow which got into the water at Aberdeen harbour has been shot after a rescue effort failed to coax it ashore.
Streets in Wales are blighted by discarded cigarette butts with 86 of roads strewn with smoking related litter a charity survey shows.
It is officially a regeneration area and dyke house in Hartlepool has newly built smart houses but they are in the minority.
Wales flyhalf Dan Biggar says he is learning to cope with the pressure of wearing the famous number 10 jersey.

Table A1: Examples of gold selections in summaries taken from both dowstream tasks with Clozer. Highlighted in yellow is the gold answer chosen according to the first definition of abstractness, *imperceptibility*, and in blue the answer according to the second definition, *non-specificity* (for hypernyms), in each example.

For task 1 (ReCAM 1), abstractness follows the definition of imperceptibility, meaning any concept that can't be perceived directly in the physical world according to a psycholinguistic database (Coltheart, 1981). Task 2 (ReCAM 2) defines abstractness as non-specificity, representing nouns and verbs relatively high in a hypernym hierarchy (Changizi, 2008). Examples of the difference between both are illustrated in Table A1.

As discussed in §5.3, the abstract concepts chosen for ReCAM 1 are intuitively harder to define compared to the concepts for ReCAM 2, even for humans (**pride, suspicion** vs **picked, following**). However, this also shows that without being given any rules, our Clozer still manages to grasp the underlying mechanics originally chosen to extract the abstract words in both tasks.

We refer to the original work (Zheng et al., 2021) on building the ReCAM dataset for more details on the reason why those two definitions of abstractness have been chosen.