SCDE: Sentence Cloze Dataset with High Quality Distractors From Examinations

Xiang Kong[*], **Varun Gangal**[*], **Eduard Hovy**
Language Technologies Institute
Carnegie Mellon University
{xiangk,vgangal,hovy}@cs.cmu.edu

Abstract

We introduce SCDE, a dataset to evaluate the performance of computational models through sentence prediction. SCDE is a human-created *sentence* cloze dataset, collected from public school English examinations. Our task requires a model to fill up multiple blanks in a passage from a shared candidate set with distractors designed by English teachers. Experimental results demonstrate that this task requires the use of *non-local*, *discourse-level* context beyond the immediate sentence neighborhood. The blanks require *joint* solving and significantly impair each other's context. Furthermore, through ablations, we show that the distractors are of high quality and make the task more challenging. Our experiments show that there is a significant performance gap between advanced models (72%) and humans (87%), encouraging future models to bridge this gap.[1] [2]

1 Introduction

Cloze questions were first proposed by Taylor (1953) as a readability test, motivated by Gestalt psychology. They become an efficient way of testing reading for public exams, overtaking the dominant paradigm of subjective questions (Fotos, 1991; Jonz, 1991). Cloze datasets (Zweig and Burges, 2011; Hermann et al., 2015; Hill et al., 2015; Paperno et al., 2016; Onishi et al., 2016; Xie et al., 2018) became prevalent as question-answering (QA) benchmarks since they are convenient either to be generated automatically or by annotators. These datasets could be split into two clear types:

1. Where the context is a complete text, and there is an explicit question posed which is a statement with a cloze gap. The answer is either generated freely or is a span from the context, e.g. Children's Books Test (CBT) (Hill et al., 2015).

2. Where the context itself comes with cloze gaps. There is no explicit question. The answer is generated freely or chosen from a set of candidates, e.g. CLOTH (Xie et al., 2018).

Herein, we focus on the 2nd category. A common property of these datasets is that they have gaps at the level of words, entities or short syntactic spans. The entity and span-based clozes may sometimes be multi-token, but they do not extend beyond a few tokens. Nevertheless, none of these datasets have cloze gaps at the level of *full sentences*. Since many syntactic and semantic cues are present in the same sentence, this makes the gap easier to fill compared to the sentence level cloze case where models would have to rely on "discourse" cues beyond the same sentence.

Besides lack of intra-sentence cues, sentence-level cloze may require comparing candidates of very different lengths. For instance, the example in Table 1 has a standard deviation of 7.6 with candidate lengths between 3 to 25. A model that only represents words well may not get comparable probabilities at sentence level for very different sentence lengths. Therefore, robust sentence representation models are also required to solve this question. In this paper, we present SCDE, a dataset of sentence-level cloze questions sourced from public school examinations. Each dataset example consists of a passage with multiple sentence-level blanks and a shared set of candidates. Besides the right answer to each cloze in the passage, the candidate set also contains ones which don't answer any cloze, a.k.a., *distractors*. Both cloze positions and distractors are authored by teachers who design the public school examinations carefully. §3.2 explains our data collection. A representative example from SCDE is shown in Table 1.

[*] Equal Contribution
[1] Data: vgtomahawk.github.io/sced.html
[2] Code: https://github.com/shawnkx/SCDE

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5668–5683
July 5 - 10, 2020.

Passage:
A student's life is never easy. And it is even more difficult if you will have to complete your study in a foreign land. _____1 The following are some basic things you need to do before even seizing that passport and boarding on the plane. Knowing the country. You shouldn't bother researching the country's hottest tourist spots or historical places. You won't go there as a tourist, but as a student. _____2 In addition, read about their laws. You surely don't want to face legal problems, especially if you're away from home. _____3 Don't expect that you can graduate abroad without knowing even the basics of the language. Before leaving your home country, take online lessons to at least master some of their words and sentences. This will be useful in living and studying there. Doing this will also prepare you in communicating with those who can't speak English. Preparing for other needs. Check the conversion of your money to their local currency. _____4. The Internet of your intended school will be very helpful in findings an apartment and helping you understand local currency. Remember, you're not only carrying your own reputation but your country's reputation as well. If you act foolishly, people there might think that all of your countrymen are foolish as well. _____5

Candidates:
A. Studying their language.
B. That would surely be a very bad start for your study abroad program.
C. Going with their trends will keep it from being too obvious that you're a foreigner.
D. Set up your bank account so you can use it there, get an insurance, and find an apartment.
E. It'll be helpful to read the most important points in their history and to read up on their culture.
F. A lot of preparations are needed so you can be sure to go back home with a diploma and a bright future waiting for you.
G. Packing your clothes.

Answers with Reasoning Type:
1→F (*Summary*) , 2→E (*Inference*) , 3→A (*Paraphrase*) , 4→D (*WordMatch*), 5→B (*Inference*) (C and G are distractors)

Discussion:
Blank 3 is the easiest to solve, since "Studying their language" is a near-paraphrase of "Knowing even the basics of the language". Blank 2 needs to be reasoned out by *Inference* - specifically E can be inferred from the previous sentence. Note however that C is also a possible inference from the previous sentence - it is only after reading the entire context, which seems to be about learning various aspects of a country, that E seems to fit better. Blank 1 needs *Summary* $\rightarrow$ it requires understanding several later sentences and abstracting out that they all refer to *lots of preparations*. Finally, Blank 5 can be mapped to B by inferring that *people thinking all your countrymen are foolish* is *bad*, while Blank 4 is a easy *WordMatch* on *apartment* to D. The other distractor G, although topically related to preparation for going abroad, does not directly fit into any of the blank contexts

Table 1: A Representative Example from SCDE.

Another salient aspect of our dataset is that more than 40% of blanks belong to the reasoning category *"Inference"* (more on this in §3.3 and Table 4) which require models to compare plausibility of competing hypotheses given a premise (whether the previous or last sentence(s), or even a combination of information from the two). Filling these blanks requires the model to reason by using commonsense knowledge, factual knowledge, time gaps, etc. Some of these can be thought of as simple entailment, but more generally, many of these can be seen as requiring abductive reasoning, which is of recent interest (Bhagavatula et al., 2019; Sap et al., 2019a,b) to the NLP community. In summary, our contributions are as follows

1. We introduce the task of *sentence level cloze completion* with multiple sentence blanks and a shared candidate set with distractors.
2. We release SCDE, a sentence level cloze dataset of $\approx$ 6k passages and $\approx$ 30k blanks.
3. We estimate human performance on SCDE, and benchmark several models, including state-of-the-art contextual embeddings (Table 5). We find a significant gap of $> 15\%$ for future models to close in order to match human performance.
4. Through several ablations described in §5.6, we show that distractors designed by English teachers are of high quality and make the task more challenging.
5. We show that extra sentence level cloze questions generated automatically from an external corpus can be used to further improve model performance through data augmentation (See §5.7).

2 Related Work

Several cloze test datasets are collected to measure reading comprehension ability of machines. CNN/DailyMail (Hermann et al., 2015), an early dataset of current QA research, constructs cloze questions from article summaries, with article spans as answers. Their cloze gaps are entities and hence one or few tokens long at best. The LAMBADA dataset (Paperno et al., 2016) constructs a corpus of word level cloze gaps, such that each gap is in the last passage sentence. CBT (Hill and Simha, 2016) creates word level cloze questions by removing a word in the last sentence of every consecutive 21 sentences, with the first 20 sentences being the context. Onishi et al. (2016) curate a dataset of who-did-what type sentences with

Dataset	SL	MB	Distractors	Candidates	Position	$\|Context\|_w$
SCDE	✓	✓	Human	Shared	Anywhere	319
ROCSTORIES (2016)	✓	×	Human	-	End	25
CLOTH (2018)	×	✓	Human	Separated	Anywhere	243
LAMBADA (2016)	×	×	Exhaustive	-	End	76
CBT (2015)	×	×	Automatic	-	End	465
MRSCC (2011)	×	×	Human	-	Anywhere	20

Table 2: Comparing SCDE with previous cloze datasets. Exhaustive denotes the case where the entire vocabulary is a candidate for a word level cloze. For the single-blank case, candidate sharing is irrelevant. SL and MB mean sentence level and multi-blanks respectively. $\|Context\|_w$ is the average token length of the context.

word level blanks. The CLOTH (Xie et al., 2018) dataset collects word level cloze questions from English exams designed by teachers. MRSCC (Zweig and Burges, 2011) consists of 1,040 word level cloze questions created by human annotators.

Among recent cloze datasets, ROCStories (Mostafazadeh et al., 2016) is the closest we could find to a sentence level cloze dataset. In this task, the first 4 sentences of a 5-sentence story are provided, and the task is to choose the correct ending from a pair of candidate ending sentences. However, there are several key differences between SCDE and ROCStories. Firstly, there are multi-blanks in SCDE which are not in a fixed position and require learning cues from bidirectional contexts of varying lengths. Secondly, the endings in ROCStories have been found to contain *"annotation artifacts"* (Gururangan et al., 2018) which makes a large fraction of them predictable independent of context.

In contrast, SCDE is by design independent of artifacts, since a) given a blank, only some of our candidates are distractors, the rest being answers for other blanks. Even if one were to learn a classifier to distinguish distractors without context, the non-distractor candidates would be unresolvable without context. b) we further check how distinguishable our distractors are from non-distractors without context by training a strong classifier in this setting, as described in §5.6. The classifier obtains a reasonably low F1 score of 0.38.

In Table 2, we summarize the comparison of SCDE with cloze datasets from prior art to show its attractive aspects.

Public school examinations have been used as a data source by many earlier QA works, two prominent examples being the CLEF QA tracks (Penas et al., 2014; Rodrigo et al., 2015) and RACE (Lai et al., 2017).

3 SCDE Dataset

3.1 Sentence Cloze Test with distractors

In this task, each question consists of a passage, S, multiple sentence level blanks B, and a shared set of candidates C with distractors D, where $D \subset C$.

Problem Complexity[3] For our case, given the typical value of $|C|$ and $|B|$ being 7 and 5 respectively, the size of the answer space, $|A|$ is 2520. Thus, the chance of guessing all blanks correctly at random is only 0.04%. Moreover, there is a 48.2% probability of being entirely wrong with randomly guessing. Finally, given an answer list chosen uniformly at random, the expectation of number of distractors in the answer list is 1.4, i.e. on average, roughly one and half answers are distractors.

3.2 Data Collection and Statistics

Raw sentence cloze problems are crawled from public websites[4] which curate middle and high school English exams designed by teachers. In total, 14,062 raw passages and 68,515 blank questions are crawled from these websites and the following steps are used to clean them. Firstly, duplicate passages are removed. Secondly, when the official answer to the problems are images, two OCR toolkits[5] are employed to convert these images to text and the questions with different results from these two programs will be discarded. Finally, we remove examples which have 1) answers pointing to non-existent candidates, 2) missing or null candidates, 3) number of blanks $>$ number of candidates, 4) missing answers.

After cleaning, we obtain our SCDE dataset with 5,959 passages and 29,731 blanks. They are

[3] We defer the derivation to Appendix §1

[4] http://www.21cnjy.com/; http://5utk.ks5u.com/; http://zujuan.xkw.com/; https://www.gzenxx.com/Html/rw/.

[5] tesseract; ABBYY FineReader

Statistic	Value
Total Passages	5,959
Total Blanks	29,731
Blanks Per Passage	4.99
# Candidates Per Passage	6.79
Avg Candidates Per Blank	1.35
% Consecutive Blanks	1.28
# Words Per Passage	319.64
Vocabulary Size	48.6k
Var(Candidate Length)	19.54

Table 3: SCDE Statistics. For Consecutive Blanks, either of previous or next sentences is also a blank.

randomly split into training, validation and test sets with 4790, 511 and 658 passages respectively. The detailed statistics are presented in Table 3. We find that candidates have very different lengths and passages have long context.

3.3 In-Depth Analysis & Categorization

In order to evaluate students' mastery of a language, teachers usually design tests in a way that questions cover different aspects of a language.

Reasoning Types As illustrated with examples in Table 4, we set a four-fold categorization for the reasoning which leads to a ground truth candidate being assigned to a blank. Our reasoning type taxonomy is motivated by categorization of question types in earlier works in QA such as (Chen et al., 2016; Trischler et al., 2017)[6]. Strictly speaking, these reasoning types could co-exist. But for simplicity, we classify each blank into only one of the four.

- WORDMATCH: If the candidate has word overlap, especially of non-stopwords or infrequent phrases, with context around the blank.
- PARAPHRASE: If the candidate doesn't have an explicit word overlap with the context, but nevertheless contains words or phrases which are paraphrases of those in the context.
- INFERENCE: If the candidate is a valid hypothesis conditioned on the left context [as premise], or a necessary precondition/premise based on the right context. Note that the candidate in this case doesn't contain word overlap/paraphrases which would obviate need for inferential reasoning. The reasoning required needs not be just strict entailment (Bowman et al., 2015; Marelli et al., 2014) but could also involve abductive reasoning (Bhagavatula et al., 2019), where the candidate is just one of many likely hypothesis (premise) given the left (right) context as premise (hypothesis).
- SUMMARY: If the candidate is a summary, introduction, or conclusion of multiple sentences before or after it. In this type, unlike INFERENCE, there is no requirement to deduce and reason about new hypotheses/possibilities not present in the premise - only consolidation and rearranging of information is required.

A sample of 100 passages containing 500 blanks are manually categorized into these four categories. Examples and statistics of these four types are listed in Table 4. More than 40% blanks need inference to be solved, denoting the high difficulty of our dataset.

4 Methods

4.1 Context Length

We experiment with giving our models different amounts of context. Through this, we can explore how context length affects model performance.

1. P(N): Immediate previous (next) sentence
2. P+N: Immediate previous and next sentence
3. AP(AN): All previous (next) sentences
4. AP+AN: All previous and next sentences

AP+AN is the unablated setting, where all passage sentences are available to the model.

4.2 PMI

Before exploring deep representational approaches, we would like to find how well symbolic ones perform at this task. Starting with works such as Iyyer et al. (2015) and Arora et al. (2017), it has become convention to first benchmark simple baselines of this kind. PMI merely encodes how likely it is for a word pair to occur in consecutive sentences. It does not consider the internal sentence structures, or the relative position of the words in their respective sentence. Intuitively, it can be called a "surface-level" approach. A high performance by PMI would indicate that candidates can be matched to blanks by simple ngram statistics, without requiring sentence representation, which would make SCDE uninteresting.

[6] See Section 4.2 from both respective papers.

Type	Examples with Excerpts From Blank Context
WM (18.47%)	**1**: One day, a teacher was giving a speech to his student. He held up a *glass of water* and asked the class _____ The students answers ranged from 20g to 500g. ✓ **Candidate**: B. How heavy do you think this *glass of water* is? × **Candidate**: D. It does not matter on the weight itself. **Explanation**: WordMatch based on *glass of water*.
Para. (19.48%)	**2**: If you want time to have breakfast with your family, save some time the night before by setting out clothes, shoes and bags. _____ That's a *quarter-hour* more you could be sleeping if you bought a *coffee* maker with a timer. × **Candidate:** D. And consider setting a second alarm. × **Candidate**: F. Stick to your set bedtime and wake-up time, no matter the day. ✓ **Candidate**: G. Reconsider the *15 minutes* you spend in line at the *cafe*. **Explanation:** Need to match *15 minutes*, *quarter-hour* and *coffee, cafe*.
Infer. (41.97%)	**3**: May is a great month. _____ You can have a good time with your family. × **Candidate**: E. All the students can come to their schools. ✓ **Candidate**: F. From May 1st to 7th, we don't need to come to school. × **Candidate**: G. On May 20th, a famous sports star YaoMing comes to our school. **Explanation:** Need to infer that not coming to school → one is at home with family. Simply matching for words *May* or *school* will also match wrong candidates.
Sum. (20.08%)	**4**: How to Enjoy Life As a Teen? Are high school days equal to the "best years of your life"? Maybe not, but you can learn to make the most of your high school days _____ Whether it 's having a computer, having friends, having a good supply of food, a bed to sleep on, family that loves you, having a decent education or simply being born in this world. Be happy, and life will reward you. × **Candidate**: A. Remember that the point of life is for you to enjoy it. ✓ **Candidate**: C. Learn to appreciate small things. **Explanation**: After summarizing sentences after the blank [which describe a list of "small things"], the answer should be C. A is a strong distractor since both "enjoy" and "life" appear in the context, besides being pertinent to the topic. Indeed, our best-performing BERT-ft model chooses A as the answer.

Table 4: Blanks in a sample of 100 passages are manually categorized into four categories. For the ease of illustration, we've shown only limited context around the blanks , and 1-2 wrong candidates. WM, Para., Infer. and Sum denote WordMatch, Paraphrase, Inference and Summary respectively. More examples are in Appendix.

We estimate PMI counts (Church and Hanks, 1990) from all consecutive sentence pairs in our training split. Let f denote frequency

$$\text{PMI}(w_s, w_c) = \frac{\text{f}(w_s \in S, w_c \in C)}{\text{f}(w_s \in S)\text{f}(w_c \in C)}$$

Note that our PMI definition diverges from typical PMI since its asymmetric between w_s and w_c. Since S and C are the sets of non-terminating and non-starting sentences respectively, they overlap but aren't identical. For a pair of sentences, we find aggregate $\overline{\text{PMI}}(S, C)$ as:

$$\overline{\text{PMI}}(S, C) = \frac{1}{|C||S|} \sum_{w_c \in C} \sum_{w_s \in S} \text{PMI}(w_s, w_c)$$

This definition can be extended to all n-grams upto a certain n. We denote this by $\overline{\text{PMI}}_n$. We notice that $\overline{\text{PMI}}_n$ performance saturates after $n = 2$. Hence, in our experiments, we use $\overline{\text{PMI}}_2$.

4.3 Language Modelling

One intuitive way to solve this task is to generate the blank sentence given the context by advanced pre-trained language models (LM). Formally, suppose the blank is the ith sentence, s_i, and $s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n$ are the context. Our goal is to choose c_k from the candidate set which could maximize the joint probability $p(s_1, \ldots, s_{i-1}, c_k, s_{i+1}, \ldots, s_n)$.

Due to limited number of passages available to train a robust LM, Transformer-XL (TR.XL) Base (Dai et al., 2019), trained on WikiText-103, is employed to address this task. In order to make decoding time tractable, context length is limited to three sentences before and after the blank.

4.4 Coherence

Coherence models assign a continuous score to a sentence sequence indicative of its coherence. This score is usually unnormalized and not needed to be a probability [unlike language models].

We use the local coherence approaches implemented by the COHERE[7] framework (Smith et al., 2016). Roughly, this model works on the intuition that successive sentences exhibit regularities in syntactic patterns. Specifically, it uses n-gram patterns on linearized syntactic parses (e.g. S NP VP ...) of consecutive sentences. Once

[7] github.com/karins/CoherenceFramework

trained, this model can return a "coherence score" for any sentence sequence.

The COHERE model is first trained on all ground-truth passages from our training set, with the ground truth answers filled into the blanks. At test-time, we score each possible answer permutation using the trained COHERE model and pick the highest scoring one. Note that decoding for COHERE is by definition exhaustive, and doesn't make any assumptions by answering the blanks in a particular order.

4.5 InferSent

Conneau et al. (2017) use textual inference supervision as a signal to train a shared sentence encoder for premises and hypotheses, which can later be used as a sentence representor. We refer to this approach as INFST. Context features of a given blank and one candidate feed to two encoders in INFST respectively and classify whether this candidate is suitable to this blank. The maximum tokens of context features is set as 256. Bi-directional LSTMs with the max pooling operation are employed as our encoders. We follow the training procedure described in Conneau et al. (2017).

4.6 BERT Models

Input Representations Let c_k denotes the kth candidate. s_{-i} and s_{+i} denote the ith sentence before and after the blank respectively and $|P|$ and $|N|$ represent total number of sentences before and after the current blank respectively. Following the input convention in Devlin et al. (2018), the input sequence given various context lengths and c_k is:

1. P : [CLS]s_{-1}[SEP]c_k
2. N : [CLS]c_k[SEP]s_{+1}
3. AP : [CLS]$s_{-|P|} \dots s_{-1}$[SEP]c_k
4. AN : [CLS]c_k[SEP]$s_{+1} \dots s_{+|N|}$

To retain sentence sequentiality, the order between the context and the candidate follows that in the original passage. Furthermore, for (A)P+(A)N, we create and score one input sample for each of the context directions during prediction. The average of these two scores is taken as the final score. The maximum tokens of input is set as 256 in our experiments and only the context is truncated to meet this requirement.

BERT Next Sentence Prediction (NSP) One of the objectives in BERT pre-training stage is understanding the relationship between two sentences, which is highly correlated with our task. Therefore, we use the pre-trained BERT-Large-uncasedd with its NSP layer to predict the most appropriate candidate for each blank given its context. Specifically, BERT is employed to predict the probability of the context and the candidate being consecutive.

Type	Model	BA/PA
UNSUP	BERT	36.9/3.5
	TR.XL	32.3/2.6
FT	BERT	**71.7/29.9**
SUP	$\overline{\text{PMI}}_2$	29.8/8.4
	COHERE	23.3/1.1
	INFST	55.8/18.4
HUMAN	-	**87.1/56.3**

Table 5: Test BA/PA of various model types with EXH decoding and AP+AN context.

Finetuning BERT A wide range of NLP tasks have greatly benefited from the pre-trained BERT model. Therefore, we also finetune the pre-trained BERT-Large model on our task through sequence pair classification schema. Specifically, for each blank, its correct candidate will be labelled as 0 and the label of all other wrong candidates is 1. Batch size and number of epochs for all models are 32 and 3. We employ Adam (Kingma and Ba, 2014) as the optimizer with three different learning rates $\{1e^{-5}, 2e^{-5}, 3e^{-5}\}$. Best model selection is based on validation performance. All BERT finetuning experiments including ablation study follow this training strategy.

5 Experiments

5.1 Decoding Strategy

The decoding strategy decides how exactly we assign a candidate to each blank in the passage. Due to shared candidates, we have two strategies:

1. **INC:** Answering each blank from left to right in order. Once a blank is answered with a candidate, this candidate is unavailable for later blanks.
2. **EXH:** Exhaustively scoring all permutations of candidates to answer the blanks. The score of a permutation is simply the sum of each its

Type	Model	P	N	AP	AN	P+N	AP+AN
UNSUP	BERT+INC	33.0/2.1	34.7/4.1	29.8/2.1	15.7/0.3	34.7/2.3	27.3/1.4
	+EXH	34.2/3.2	40.2/4.7	31.5/2.6	14.7/0.0	40.2/4.7	36.9/3.5
FT	BERT+INC	44.3/6.8	48.0/9.6	50.4/9.9	56.9/16.1	61.0/20.4	66.6/25.1
	+EXH	47.2/8.5	54.2/11.2	60.0/17.5	60.0/17.5	66.5/25.2	**71.7/29.9**
SUP	$\overline{\text{PMI}}_2$+INC	23.4/1.2	24.4/1.5	16.2/0.3	17.5/0.1	26.2/1.7	17.1/0.0
	+EXH	24.7/1.5	28.2/1.5	20.6/0.9	13.3/0.0	29.7/2.6	25.2/0.6

Table 6: Test BA/PA of various model types unsupervised (UNSUP), finetuned (FT) and supervised (SUP) across varying context levels, with INC or EXH decoding.

	BERT-Un	TR.XL	BERT-ft
RemoveDt	47.4/17.2	39.7/9.1	80.9/62.0
RandomDt	44.6/12.4	36.0/6.8	77.9/50.9
Unablated	40.2/4.7	32.3/2.6	71.7/29.9

Table 7: Test BA/PA with distractor ablations on test set. RemoveDt and RandomDt represent removing and sampling distractors respectively. BERT-Un and BERT-ft denotes pre-trained and finetuned BERT.

constituent blank-candidate pairs. The highest scoring permutation is the answer.

5.2 Evaluation Metrics

We design two metrics to evaluate models. Both of these metrics are reported as percentage.

Blank accuracy (BA): The fraction of blanks answered correctly, averaged over all passages.

Passage Accuracy (PA): PA is 1 *iff* the model gets all blanks in a passage correct, and 0 otherwise. The average of PA over all passages is reported.

5.3 Human Performance

We hire annotators from AMT to both answer and label difficulty for 144 randomly chosen test examples. Annotators are restricted to be from USA/UK and have *Master* designation on AMT[8], along with $> 90\%$ HIT approval rate. On average, each annotator spends 624 seconds to answer one example. Difficulty level is chosen from {*VeryHard*, *Hard*, *Moderate*, *Easy*, *VeryEasy*}. 3.5% of annotators find the task *VeryHard*, while 8.3% find it *VeryEasy*. The largest fraction of 38.2% find it to be *Moderate*. We note that SCDE contains a larger proportion of non-easy questions (61.0%). Human performance is reported in Table 5. Annotators achieve BA of 87% which we take as the ceiling performance for models to match.

5.4 Model Performance

All models are trained with AP+AN context and decoded by EXH[9]. Results are shown in Table 5. Finetuning BERT achieves the best performance among other models, though it still lags behind human performance significantly. Unsupervised models could only solve one third of all blanks. Surprisingly, $\overline{\text{PMI}}_2$ and COHERE performs worse than the unsupervised models. We conjecture that it is difficult for COHERE, using syntactic regularities alone, to distinguish between the ground truth answer for a particular blank and another candidate which is a ground truth answer for another nearby blank. As noted, $\overline{\text{PMI}}_2$ suffers due to inability of incorporating larger context.

To explore effects of various context length and decoding strategies, models are trained with different context lengths and inferred by both decoding methods. Results are shown in Table 6.

INC vs EXH EXH is better than INC for most approaches, indicating that human created blanks are interdependent and need joint answering.

Context Length Increasing the context length, such as (P vs. AP), could significantly improve model performance, showing that this task needs discourse-level context to be answered. Furthermore, models with bidirectional context, such as (P+N), perform better than single-direction context, e.g., P, indicating that this task needs global context. Lastly, we observe that PMI-based approaches which do not explicitly encode sentences

[8]Marked by AMT based on approval %, no. approved etc.

[9]Unless stated otherwise, models decode with EXH and are trained with full context i.e AP+AN

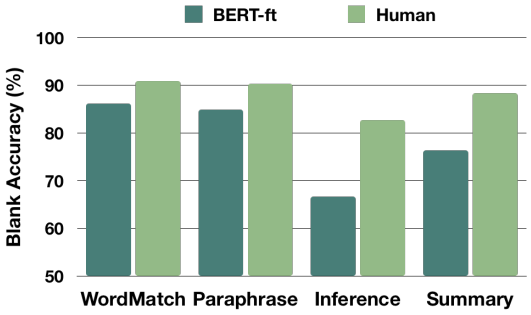

Figure 1: Test blank accuracy of BERT-ft and Human on each reasoning type category introduced in §3.3.

are unable to incorporate larger context levels, showing best performance with P+N.

5.5 BERT-ft vs. Human

BERT after finetuning (BERT-ft) can perform reasonably well (72%) but there is still a gap comparing with human performance (87%). In this section, we would like to analyze the strength and weakness of BERT-ft compared with HUMAN. Therefore, we analyze their performance across different reasoning categories on test set. From Figure 1, inference questions are the most difficult for both HUMAN and BERT-ft and questions needing WordMatch are relatively easy. Compared with human performance, BERT-ft could achieve comparable BA on WordMatch and paraphrasing problems. However, BERT-ft performs much worse on questions needing inference and summary. We also refer to some examples from Table 4.

In *Example* 4, BERT-ft prefers A but the answer is C. The reason why BERT-ft chooses A may be that "enjoy life" happens in the context, but summarizing the next sentence is necessary to achieve the correct answer. Therefore, it is necessary to improve the ability of BERT to represent meaning at the sentence level beyond representing individual words in context.

We also explore how the system performance corresponds to the human judgement of difficulty. Since evaluates rate the problems into 5 difficulty levels, we report the system BA/PA for each level in Table 8. For BA (blank-level accuracy), we see that, overall, the system accuracy decreases as difficulty increases from VeryEasy (0.75) to VeryHard (0.68). However, the decrease is not exactly monotonic (there is a small increase from VeryEasy to Easy, as also from Moderate to Hard).

We conjecture that non-monotonicity could be due to two reasons:

- Our difficulty annotations are at passage level rather than blank level. There might be some hard blanks in a passage marked overall "Easy". Conversely, there might be easy blanks in a passage marked overall "Hard".
- Since we've more examples marked with certain difficulty levels - e.g 30.5% examples are "Easy" while only 8.3% are "VeryEasy". This might make system accuracy average for levels with more examples more stable (lower sample variance), leading to some non-monotonicity (e.g for Easy and VeryEasy)

For PA (passage-level accuracy, i.e., getting all questions correct) also, we see a clear decrease as difficulty increases from VeryEasy (0.63) to VeryHard(0.2). The decrease here is sharper than BA , with only one violation of monotonicity (increase from 0.29 to 0.35 on Moderate to Hard). The sharper trend for PA supports our first point above.

Diffculty	BA	PA
Very Easy	0.75	0.63
Easy	0.78	0.45
Moderate	0.71	0.29
Hard	0.72	0.35
Very Hard	0.68	0.20

Table 8: BERT-ft performance in terms of human judgement of diffculty.

5.6 Distractor Quality

An attractive aspect of this task is distractors designed by English teachers. We verify distractor quality through the following experiments.

Model Performance w/o Distractors All distractors in the test set are removed and models are evaluated on this non-distracted test set. Results are shown in Table 7. It is clear to see that after removing these distracting candidates, models can get better scores, showing that models find it hard to exclude distractors during prediction.

Randomly Sampled Distractors After removing human-created distractors, we further randomly sample sentences from other passages as new distractors. To mitigate sampling variance, we run this experiment with 8 seeds and report the

Model	Uni.	$\overline{\text{PMI}}_2$	BERT-ft	HUMAN
DE	1.429	1.204	0.661	**0.375**

Table 9: Distractor error on test set of different models. Uni. denotes the uniform model.

Training Strategy	PA	BA	DE
Q_A	65.2	26.1	0.792
Q_H	71.7	29.9	0.661
Q_A ; Q_H	74.2	33.9	**0.624**
$Q_A + Q_H$	**74.5**	**34.3**	0.637

Table 10: Test performance of models with Q_A and Q_H.

averaged score in Table 7. Comparing with distractors designed by teachers, models could discern these distractors more easily.

Annotation artifacts of distractors Annotation artifacts (Gururangan et al., 2018) occurs in many datasets created by human annotators. A potential artifact type for our task is whether we could detect distractors without passages. Therefore, we finetune BERT-Large as a binary classifier, the input of which is just distractors and other correct candidates. With this model, we could only obtain 38% F1 score on the test set, showing that it is difficult to filter distractors out without any context.

Distractor Error (DE) We define DE as the number of predicted answers per passage which are actually distractors. Through DE, we measure a model's ability to exclude distractors during prediction. Results are shown in Table 9. HUMAN has the lowest DE and BERT-ft could discern distractors to some extent. However, DE of $\overline{\text{PMI}}_2$ is more than 1, meaning that on average, there is atleast one distractor in the predicted answer list.

In summary, distractors created by teachers are high quality and increase task difficulty.

5.7 Automatically Generated Sentence Cloze Questions

To explore automatic generation of examples for the task, we construct sentence cloze questions by randomly choosing five sentences in a passage as blanks. We defer automatically generating distractors to future work since non-trivial distractor generation is a hard problem in itself. Specifically, we extract all passages from RACE (Lai et al., 2017) (which is also from exams) and filter out passages which have less than 10 sentences or more than 30 sentences. While choosing blank positions, we prevent three or more blanks consecutive to each other in generated questions. Finally, 16,706 examples are obtained automatically. Here, questions generated automatically and collected from examinations are called Q_A and Q_H respectively.

We leverage Q_A in three ways: 1). train models only on Q_A , 2) first train models on Q_A and finetune models on Q_H, i.e., Q_A ; Q_H, 3) train models on the concatenation of Q_A and Q_H, i.e., $Q_A + Q_H$. BERT-Large is finetuned through these ways and results are shown in Table 10. The model trained only on Q_A has worst performance and we attribute this to the difficulty of distinguishing distractors without seeing them during training. Therefore, this model has the highest DE. However, models trained on Q_H and Q_A could achieve better performance. We conjecture this is because Q_A assists the model to have better generalization.

6 Conclusion

We introduce SCDE, a sentence cloze dataset with high quality distractors carefully designed by English teachers. SCDE requires use of discourse-level context and different reasoning types. More importantly, the high quality distractors make this task more challenging. Human performance is found to exceed advanced contextual embedding and language models by a significant margin. Through SCDE, we aim to encourage the development of more advanced language understanding models.

Acknowledgements

We thank Qizhe Xie, Hiroaki Hayashi and the 3 anonymous reviewers for valuable comments.

References

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. *ICLR*.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Danqi Chen, Jason Bolton, and Christopher D Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Zihang Dai, Zhilin Yang, Yiming Yang, William W Cohen, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

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*.

Sandra S Fotos. 1991. The cloze test as an integrative measure of efl proficiency: A substitute for essays on college entrance examinations? *Language learning*, 41(3):313–336.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 107–112.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

Jennifer Hill and Rahul Simha. 2016. Automatic generation of context-based fill-in-the-blank exercises using co-occurrence likelihoods and google n-grams. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 23–30.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1681–1691.

Jon Jonz. 1991. Cloze item types and second language comprehension. *Language testing*, 8(1):1–22.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 1–8.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.

Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2016. Who did what: A large-scale person-centered cloze dataset. *arXiv preprint arXiv:1608.05457*.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The lambada dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*.

Anselmo Penas, Yusuke Miyao, Alvaro Rodrigo, Eduard H Hovy, and Noriko Kando. 2014. Overview of CLEF QA Entrance Exams Task 2014. In *CLEF (Working Notes)*, pages 1194–1200.

Alvaro Rodrigo, Anselmo Penas, Yusuke Miyao, Eduard H Hovy, and Noriko Kando. 2015. Overview of CLEF QA Entrance Exams Task 2015. In *CLEF (Working Notes)*.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019b. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Karin Sim Smith, Wilker Aziz, and Lucia Specia. 2016. Cohere: A toolkit for local coherence. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4111–4114.

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy. 2018. Large-scale cloze test dataset created by teachers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2344–2356.

Geoffrey Zweig and Christopher JC Burges. 2011. The microsoft research sentence completion challenge. *Microsoft Research, Redmond, WA, USA, Tech. Rep. MSR-TR-2011-129*.

A Problem Complexity

With $|B| = 5$ blanks and $|C| = 7$ candidates, the size of answer space, $|A|$, is number of permutations $|B|$ objects taken $|C|$ at a time, i.e., $\mathrm{P}(7, 5) = 2520$. Therefore, the probability of answering all blanks correctly is $\frac{1}{2520} = 0.03\%$

What are the chances of getting answers partially correct? What are the chances of getting answers partially correct? If we have the same number of candidates as blanks, this is equivalent to $|B|! - D_{|B|}$, where $D_{|B|}$ is the number of derangements[10] of $|B|$ elements. In the presence of more candidates than blanks i.e distractors, this expression becomes more involved to derive. Therefore, here, we enumerate all the permutation of answer lists given a correct answer. With $|C| = 7$ and $|B| = 5$, $\zeta(|C|, |B|) = 51.8\%$. In other words, there is a 48.2% probability of being entirely wrong with a randomly chosen set of answers to each blank in the passage.

What are the chances of getting distractors as predicted answers? For the expectation of number of distractors choosing by uniform model, it should be $\mathbb{E}[DE]$, where DE denotes distractors errors.

$$\sum_{d=0}^{2} p(DE = d) \times d \tag{1}$$

where $p(DE = d)$ denotes the probability of d predicated answers are distractors. Since there are two distractors in candidates, the maximum of d is 2. Furthermore, $p(DE = 1)$ is

$$\mathrm{P}(5, 4)\mathrm{C}(5, 4)\mathrm{C}(2, 1)/|A| = 0.476 \tag{2}$$

and $p(DE = 2)$ is

$$\mathrm{P}(5, 3)\mathrm{C}(5, 3)\mathrm{A}(2, 2)/|A| = 0.476 \tag{3}$$

where $\mathrm{C}(\cdot, \cdot)$ and $\mathrm{P}(\cdot, \cdot)$ is combination and permutation respectively. Therefore, the expectation of number of distractors is 1.429.

B Additional Experiment Specifications

Specific BERT Model Used

We use uncased BERT models for all our experiments. We use the BERT models trained by the canonical pytorch implementation of Wolf et al. (2019).

[10] en.wikipedia.org/wiki/Derangement

C More examples

We show more examples belonging to different reasoning categories in Table 11. Also, some completed questions with strong distractors, multi-blank logic and diverse reasoning types are shown in Table 12, 13 and 14.

Reasoning	Examples with Excerpts From Blank Context
WM (18.47%)	**1**: One day, a teacher was giving a speech to his student. He held up a *glass of water* and asked the class. _____ The students answers ranged from 20g to 500g. ✓ **Candidate**: B. How heavy do you think this *glass of water* is? × **Candidate**: D. It does not matter on the weight itself. **Explanation**: Match based on *glass of water* **2**: Begin the sleep adjustment for your school schedule as *early* as possible. _____ But if you feel you will need some extra time to adjust, *start earlier*. ✓ **Candidate**: C. *Starting* a few days *early* will be enough. × **Candidate**: A. Relax before you go to bed. **Explanation**: Match based on *early*, *start*
Para. (19.47%)	**3**: If you want time to have breakfast with your family, save some time the night before by setting out clothes, shoes, and bags. _____ That's a *quarter-hour* more you could be sleeping if you bought a *coffee* maker with a timer. ✓ **Candidate**: G. Reconsider the *15 minutes* you spend in line at the *cafe*. × **Candidate**: F. Stick to your set bedtime and wake-up time, no matter the day. × **Candidate:** D. And consider setting a second alarm **Explanation:** Need to match *15 minutes*, *quarter-hour* and *coffee, cafe* **4**: Riding a London subway, a person from China will notice one major difference: In London, commuters do not look at each other. _____ That's not rudeness- people are just too busy to bother *looking*. ✓ **Candidate**: E. In fact, *eye contact* is avoided at all times. × **Candidate:** F. Apple must earn a fortune from London commuters. × **Candidate:** G. Modern Londoner are fancy victims. **Explanation:** Need to match *looking* and *eye contact*
Infer. (41.16%)	**5**: May is a great month. _____ You can have a good time with your family. ✓ **Candidate**: F. From May 1st to 7th, we don't need to come to school. × **Candidate**: G. On May 20th, a famous sports star YaoMing comes to our school. × **Candidate**: E. All the students can come to their schools. **Explanation:** Need to infer that not coming to school → one is at home with family. Simply matching for words *May* or *school* will also match wrong candidates. **6**: The Colosseum in Rome was built during the time of the Roman Empire, in the first century AD. _____. It is a popular tourist attraction today. ✓ **Candidate**: D. It could seat 50K people, who went to see fights between animals and people. × **Candidate:** B. The country used to depend on agriculture. × **Candidate:** C. Mountains cover about three-fourths of the country. **Explanation:** World knowledge that *Colosseum* or *-eum* suffix relates to building with seating facility. Also coreference with the *It* in *It is a popular . . .* **7**: American students usually get to school at about 8 : 30 in the morning. _____ In class, American students can sit in their seats when they answer teachers' questions. ✓ **Candidate**: B. School starts at 9:00 a.m. × **Candidate:** D. Then they take part in different kinds of after-school activities. **Explanation:** Requires inference about time. Activity starts at 9 after participants get there before.
Sum. (20.08%)	**8**: Around water, adults should watch children at all times to make sure they are safe. Those who don't know how to swim should wear life jackets. But by themselves they are not enough, so an adult should always be present. If you have to rescue a child from drowning, a few seconds can make a big difference. Make sure you have a friend with you whenever you swim. _____. That person can make sure you get help. Drink a lot water. The sun's heat and the physical activity may make you sweat more than you realize. By following these simple rules, you can make sure your swim time is safe as well as fun. _____ ✓ **Candidate**: B. Now get out there, and enjoy the water. × **Candidate**: D. Make sure everyone in your family swim well. **Explanation**: B is a good conclusion pertinent to the content of the passage. **9**: _____. Whenever you are worried, write down the questions that make you worry. And write out all the various steps you could take and then the probable consequences of each step. For example, "What am I worrying about?", What can I do about it? Here is what I'm going to do about it. After carefully weighing all the facts, you can calmly come to a decision. ✓ **Candidate**: A. Analyze the facts. × **Candidate**: C. Decide how much anxiety a thing may be worth. **Explanation**: A is a more appropriate option to summarize its succeeding context.

Table 11: More examples of reasoning categories.

Dear David _____1 After I had spent a week with my English family, I slowly began to understand their English a little better. _____2 Students in my group are from different cities of Britain and their dialects are different too! Some of their accents are quite strong and they also have their own words and expressions. _____3 Before I came to England I had thought that fish and chips were eaten every day. That's quite wrong! I get rather annoyed now when I hear all the foolish words about typical English food. I had expected to see "London fog". Do you remember our texts about it ? We had no idea that most of this "thick fog" disappeared many years ago when people stopped using coal in their homes. But the idea to speak about weather was very helpful. _____4 On the other hand , habits are different . People tell me what is typical British here in London is not always typical in Wales or Scotland. _____5 But what is ordinary for all British is that they follow traditions. Probably Britain has more living signs of its past than many other countries. And people have always been proud of having ancient buildings in capitals, big cities and the countryside. I will tell you more about Britain in my other letters. Love from Britain.

Candidates:
A. But it's not the language that's different and surprising.
B. Thanks for your nice letter.
C. I have difficulty in understanding my classmates.
D. The family I live with are friendly.
E. It 's very different from what I learned at school.
F. Local habits and traditions are not the same as what we knew.
G. The weather in London is really changeable.
Answers: 1→B , 2→E, 3→A , 4→G, 5→F (C and D are distractors)

Discussion: C is a strong distractor - not only does it have strong word overlap with the contexts of many blanks - it also has words which can make it rank high in terms of the possible inferences (dialects are different implies difficulty in understanding. Though not as strong as C, D also has a key word matching and is similar in content to the topic.

How to Enjoy Life As a Teen. Are high school days equal to the "best years of your life"? Maybe not, but you can learn to make the most of your high school days. _____1 Whether it's having a computer, having friends, having a good supply of food, a bed to sleep on, family that loves you, having a decent education or simply being born in this world. Be happy, and life will reward you. Remember that these are the last few years you will be able to enjoy yourself without having to worry about the responsibility of an adult, but make sure you prepare yourself for when you do become one. Choose your friends wisely. Unlike what many articles state, you don't have to be popular and have a gazillion friends to be happy. _____2 Try to have friends that like you who you are, not just because you are wearing a certain brand of shoes or something like that. These are people who shop at the same store as you; not someone who will sympathize with you when your dog dies. _____3 Participating in clubs, activities, and sports increases your chances of meeting new friends. While you only need 4 or 5 close friends, that doesn't mean you shouldn't try to meet new people. Participating gives you something to do instead of sitting bored at home and wallowing in self-pity. You can pursue interests you enjoy. Video games, for example, are good if you're the type who can get into that kind of thing. Use your "hobby time" either to gain practical skills for college apps, job resumes, and scholarships or get into something else in the creative field like painting or dance. _____4 Work at a job you can enjoy. Working is a great way to gain experience and to meet other people. When you do get out of college, interviewing companies will look at your prior work experience. _____5 If you can't find work, especially in this hard economic time, volunteer or make your own job.

Candidates:
A.Remember that the point of life is for you to enjoy it.
B. In fact, many of the "friends" you have when you are popular are not true friends.
C. Learn to appreciate small things.
D. Be sociable.
E. This will look great on your resume.
F. This is the time to start developing passions.
G. You should also find a hobby that is meaningful or practical.
Answers: 1→C , 2→B, 3→D , 4→F, 5→E (A and G are distractors)

Discussion: Both A and G are strong distractors especially for _____4. Both of them overlap on key words, and do fit in the local context, though they are less coherent w.r.t F (which doesn't have any overlapping words) when placed in the broader narrative.

Table 12: Examples with strong distractors

The demand for ways to improve memory is higher in students than it is in adults. Students often come across new knowledge in different areas that they need to store for exams. _____1 Here are three effective ways to improve your memory as a student. _____2 Research shows that learning activities that take more than two hours without a break are less productive when compared to those that take one hour or 30 minutes. Students are likely to remember things they learn over a short period of time. Make sure you take breaks between learning sessions to help improve your memory. Try to relax. Relaxing should be an essential part of your learning process. Scientists have proven that stronger and lasting memories can be achieved when a person relaxes. _____3 Deep breathing is one of the most popular relaxation techniques. Establish a quiet environment and find a comfortable position. Then go through a deep breathing process for at least 15 minutes. Train the brain Students should give their brains a workout in order to improve their memory. At times the brain needs the right stimulation to keep growing and developing. You need to come up with a brain boosting activity that is suitable for you. _____4 Write a short story and then try to use seven to nine words to describe it. You can also do games and puzzles to help improve your memory. _____5 The techniques discussed above will help you to improve your memory significantly.

Candidates:
A. Distribute learning.
B. Enrich learning activities.
C. Some students suffer with memory problems.
D. Like a muscle memory can stretch and grow with a workout.
E. For instance you can prepare a list of items and try to memorize them.
F. You need to use different relaxation techniques in order to improve your memory.
G. In summary a good memory is an important advantage to any student who wants to improve his or her grades.

Answers: 1→C, 2→A, 3→F , 4→E, 5→G (B and D are distractors)

Discussion: The candidate F can actually go into three possible blanks and fit well into their context - Blanks 1, 3 and 5. This can be seen from the several overlapping phrases/paraphrases F shares with all three, as shown by the three colors (one per concept). However, G (which starts with the phrase *In summary,* can only fit into Blank 5. A is also difficult to place in any blank other than Blank 1. Hence , candidate F has to be placed into Blank 3.

Table 13: Examples which require multi-blank logic

A student's life is never easy. And it is even more difficult if you will have to complete your study in a foreign land. _____1 The following are some basic things you need to do before even seizing that passport and boarding on the plane. Knowing the country. You shouldn't bother researching the country's hottest tourist spots or historical places. You won't go there as a tourist, but as a student. _____2 In addition, read about their laws. You surely don't want to face legal problems, especially if you're away from home. _____3 Don't expect that you can graduate abroad without knowing even the basics of the language. Before leaving your home country, take online lessons to at least master some of their words and sentences. This will be useful in living and studying there. Doing this will also prepare you in communicating with those who can't speak English. Preparing for other needs. Check the conversion of your money to their local currency. _____4 The Internet of your intended school will be very helpful in findings an apartment and helping you understand local currency. Remember, you're not only carrying your own reputation but your country's reputation as well. If you act foolishly, people there might think that all of your countrymen are foolish as well. _____5

Candidates:
A. Studying their language.
B. That would surely be a very bad start for your study abroad program.
C. Going with their trends will keep it from being too obvious that you're a foreigner.
D. Set up your bank account so you can use it there , get an insurance , and find an apartment.
E. It'll be helpful to read the most important points in their history and to read up on their culture.
F. A lot of preparations are needed so you can be sure to go back home with a diploma and a bright future waiting for you.
G. Packing your clothes.

Answers with Reasoning Type:
1→F (*Summary*), 2→E (*Inference*), 3→A (*Paraphrase*), 4→D (*WordMatch*), 5→B (*Inference*) (C and G are distractors)

Discussion: Blank 3 is the easiest to solve, since *Studying their language* is a near-paraphrase of *Knowing even the basics of the language*. Blank 2 needs to be reasoned out by *Inference* - specifically E can be inferred from the previous sentence. Note however that C is also a possible inference from the previous sentence - it is only after reading the entire context, which seems to be about learning various aspects of a country, that E seems to fit better. Blank 1 needs to be reasoned out by *Summary* → it requires understanding several later sentences and abstracting out that they all refer to *lots of preparations*. Finally, Blank 5 can be mapped to B by inferring that *people thinking all your countrymen are foolish* is *bad*, while Blank 4 is a easy *WordMatch* on *apartment* to D.

Latest news and comment on Street art from guardian.co.uk... _____1 You can find it on buildings sidewalks street signs and trash cans from Tokyo to Paris from Moscow to Cape Town. Street art has become a global culture and even art museums and galleries are collecting the works of street artist. Street art started out very secretly because it was illegal to paint on public and private property without permission. _____2 Some think it is a crime and others think it is a very beautiful new form of culture. Art experts claim that the street art movement began in New York in the 1960s. Young adults painted words and other images on the walls and trains. This colorful style of writing became known as graffiti whose art showed that young people wanted to rebel against society. Street artists do their work for different reasons. _____3 They choose street art because it is closer to the people. Some artists try to express their political opinion in their work. Others like to do things that are forbidden and hope they don't caught. Advertising companies also use street art in their ads because it gives people the impressions of youth and energy. _____4 Artists can show their pictures to an audience all over the world. Many city residents however say that seeing a picture on the Internet is never as good as seeing it alive. _____5. There it will continue to change and grow

Candidates:
A. Street art is a very popular form of art that is spreading quickly all over the world.
B. Today the Internet has a big influence on street art.
C. With the development of science and technology different art styles come into the Internet.
D. The street art movement lives with the energy and life of a big city.
E. People often have different opinions about street art.
F. Street art used to be illegal but now has become popular.
G. Some of them do not like artists who make so much money in galleries and museums.

Answers with Reasoning Type:
1→A (*Summary*), 2→E (*Inference*), 3→G (*Inference*), 4→B (*Inference*), 5→D (*Inference*) (C and F are distractors)

Discussion: Blank 1 requires an answer which makes an overall broad statement to introduce the topic. Working backwards, this requires summarizing or finding a broad topic given the latter sentences.

Table 14: Representative examples with diverse reasoning types