No Title Given

No Author Given

No Institute Given

Supplementary Material

In this report we show a few answer prediction examples from the test set, and present an analysis on the performance of our QA systems (BERT_{CR}, BERT_{NMT} and $\text{BERT}_{\text{MULT3}}$) and $\text{BERT}_{\text{BASE}}$ in the QA tasks. The tables show questions and the paragraphs containing the answers of the questions.

Table 1. Answer Predictions by $\text{BERT}_{\text{BASE}}$ and BERT_{CR}.

Paragraph	The Panthers finished the regular season with a 15–1 record, and quarterback Cam Newton was named the NFL Most Valuable Player (MVP). They defeated the Arizona Cardinals 49–15 in the NFC Championship Game and advanced to their **second** Super Bowl appearance since the franchise was founded in 1995. The Broncos finished the regular season with a 12–4 record, and denied the New England Patriots a chance to defend their title from Super Bowl XLIX by defeating them 20–18 in the AFC Championship Game. They joined the Patriots, Dallas Cowboys, and Pittsburgh Steelers as one of four teams that have made **eight** appearances in the Super Bowl.
Question	How many times have the Panthers been in the Super Bowl?
$\text{BERT}_{\text{BASE}}$	eight
BERT_{CR}	second

Explanation: this example shows a question and the paragraph containing the answer of the question. We see that the answer predicted by $\text{BERT}_{\text{BASE}}$ is incorrect (i.e. eight), but our CR-based QA system, BERT_{CR}, predicted the answer correctly (i.e. second). As can be seen from the paragraph, the main difficulty in answering this question is owing to the fact that there are many team names (i.e. proper nouns) present in the text, and the named entity 'Super Bowl' appears in two different positions in the text. In Figure 1, we show how pronouns ('they') are related to the original referents ('Panthers', 'Broncos'). In other words, Figure 1 shows a portion of the output of the CR tool for the paragraph. As far as the training (i.e. fine-turning) and prediction of BERT_{CR} are concerned, all such entities were substituted with the original referring entities in the paragraphs. We conjecture that this has helped BERT_{CR} to predict the right answer.

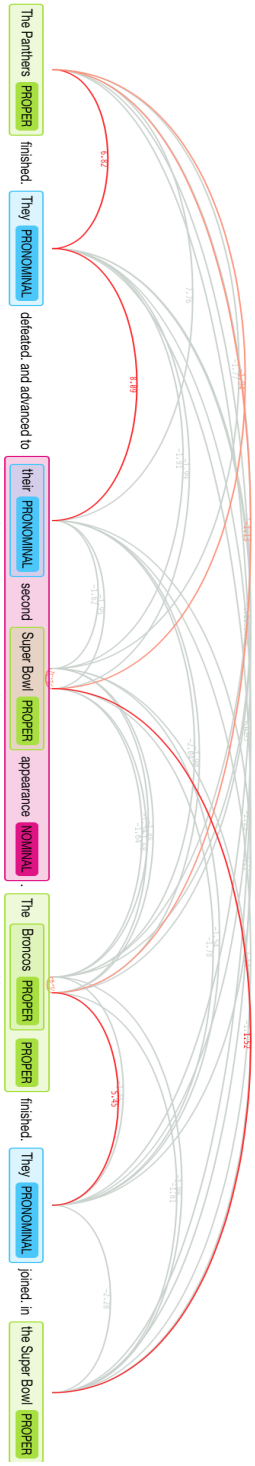

Fig. 1. Mapping between pronominal entities and referents.

Table 2. Answer Predictions by BERT_{CR} and BERT_{NMT}

Paragraph	In 1882, Tesla began working for the Continental Edison Company in France, designing and making improvements to electrical equipment. In June 1884, he relocated to **New York City**, where he was hired by Thomas Edison to work at his **Edison Machine Works** on Manhattan's lower east side. Tesla's work for Edison began with simple electrical engineering and quickly progressed to solving more difficult problems.
Question	Where did Tesla begin working in 1984?
$\text{BERT}_{\text{BASE}}$	New York City
BERT_{NMT}	Edison Machine Works

Explanation: we see from that the answer predicted by $\text{BERT}_{\text{BASE}}$ is wrong, and BERT_{NMT} predicts the answer correctly. When we look at the contexts of the question and paragraph, we can see that the main challenge of the QA model would be to disambiguate the predictions, i.e. choosing the company where Tesla was hired for working or the place where Tesla lived. As discussed in our paper, we obtain many variants (e.g. 12 in this case) of a question using our MT-based QE strategy. Here, we report some of the variants of the original question '*Where did Tesla begin working in 1984?*' that we obtained from the MT system: '*Where did Tesla work in 1984?*', '*Where did Tesla start working in 1984?*', ' *Where did Tesla start job in 1984?* '. The additional contexts in the form of alternative question sequences of the original question provides more reasoning knowledge to BERT. We conjecture that this helped BERT_{NMT} to predict the correct answer.

Table 3. Answer Predictions by $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{MULT3}}$.

Paragraph	A modern example of school discipline in North America and Western Europe relies upon the idea of an assertive teacher who is prepared to impose their will upon a class. Positive reinforcement is balanced with immediate and fair punishment for misbehavior and firm, clear boundaries define what is appropriate and **inappropriate behavior**. Teachers are expected to respect their students; **sarcasm and attempts to humiliate pupils** are seen as falling outside of what constitutes proper discipline.
Question	What is not considered appropriate discipline?
$\text{BERT}_{\text{BASE}}$	inappropriate behavior
$\text{BERT}_{\text{MULT3}}$	sarcasm and attempts to humiliate pupils

Paragraph	In **1564** a group of Norman Huguenots under the leadership of Jean Ribault established the small colony of Fort Caroline on the banks of the St. Johns River in what is today Jacksonville, Florida. The effort was the first at any permanent European settlement in the present-day continental United States, but survived only a short time. At September **1565** French naval attack against the new Spanish colony at St. Augustine failed when its ships were hit by a hurricane on their way to the Spanish encampment at Fort Matanzas. Hundreds of French soldiers were stranded and surrendered to the numerically inferior Spanish forces led by Pedro Menendez. Menendez proceeded to massacre the defenceless Huguenots, after which he wiped out the Fort Caroline garrison.
Question	When was the colony destroyed?
$\text{BERT}_{\text{BASE}}$	1564
$\text{BERT}_{\text{MULT3}}$	1565

Explanation: $\text{BERT}_{\text{BASE}}$ failed to correctly predict the answers for the both questions, which, however, were correctly predicted by $\text{BERT}_{\text{MULT3}}$. As far as the first question is concerned, as in Table 2, the main challenge for the QA model was again to disambiguate the contexts of the question and paragraph text. $\text{BERT}_{\text{MULT3}}$ was trained on the sequences made of expanded input questions (cf. QE techniques in the paper) and modified paragraphs (references to the original referring entities substituted with the original referring entities) For the first question we obtained the relevant synonyms of the good term 'appropriate' as 'proper', 'suitable', 'relevant'. In addition to this, the NMT system provided 12 similar variants of the question, one of which is "What is not considered proper discipline?". This additional contextual information of the original question essentially provides more reasoning knowledge to the model, which is probably the reason why $\text{BERT}_{\text{MULT3}}$ found the right answer for the first question. In the secodn case, we found that the WordNet-based QE technique provided the relevant lexical variations of the good term 'destroyed': 'ruin', 'destruct', 'demolish' and the NMT-based QE technique provided the alternative question sequences for the original question (one of the translations is "When was the colony wiped out?"). In addition to these, CR included more direct knowledge in the paragraph (e.g. the referring word 'he' substituted by the referent 'Menendez'). These processes may lead $\text{BERT}_{\text{MULT3}}$ to find the right answer. In comparison, $\text{BERT}_{\text{BASE}}$ failed to find the correct answer.