DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs

ACL 2019

Dheeru Dua♣, Yizhong Wang♦∗ , Pradeep Dasigi♥,
Gabriel Stanovsky♥+, Sameer Singh♣, and Matt Gardner♠

♣University of California, Irvine, USA ♦Peking University, Beijing, China
♥Allen Institute for Artificial Intelligence, Seattle, Washington, USA
♠Allen Institute for Artificial Intelligence, Irvine, California, USA
+University of Washington, Seattle, Washington, USA

Part 1.

Introduction

- **The task of reading comprehension**
  - Systems must understand a single passage of text well enough to answer arbitrary questions about it, has seen significant progress in the last few years
  - So much that the most popular datasets available for this task have been solved
  - Introduce a substantially more challenging English reading comprehension dataset
    - It aimed at pushing the field towards more comprehensive analysis of paragraphs of text

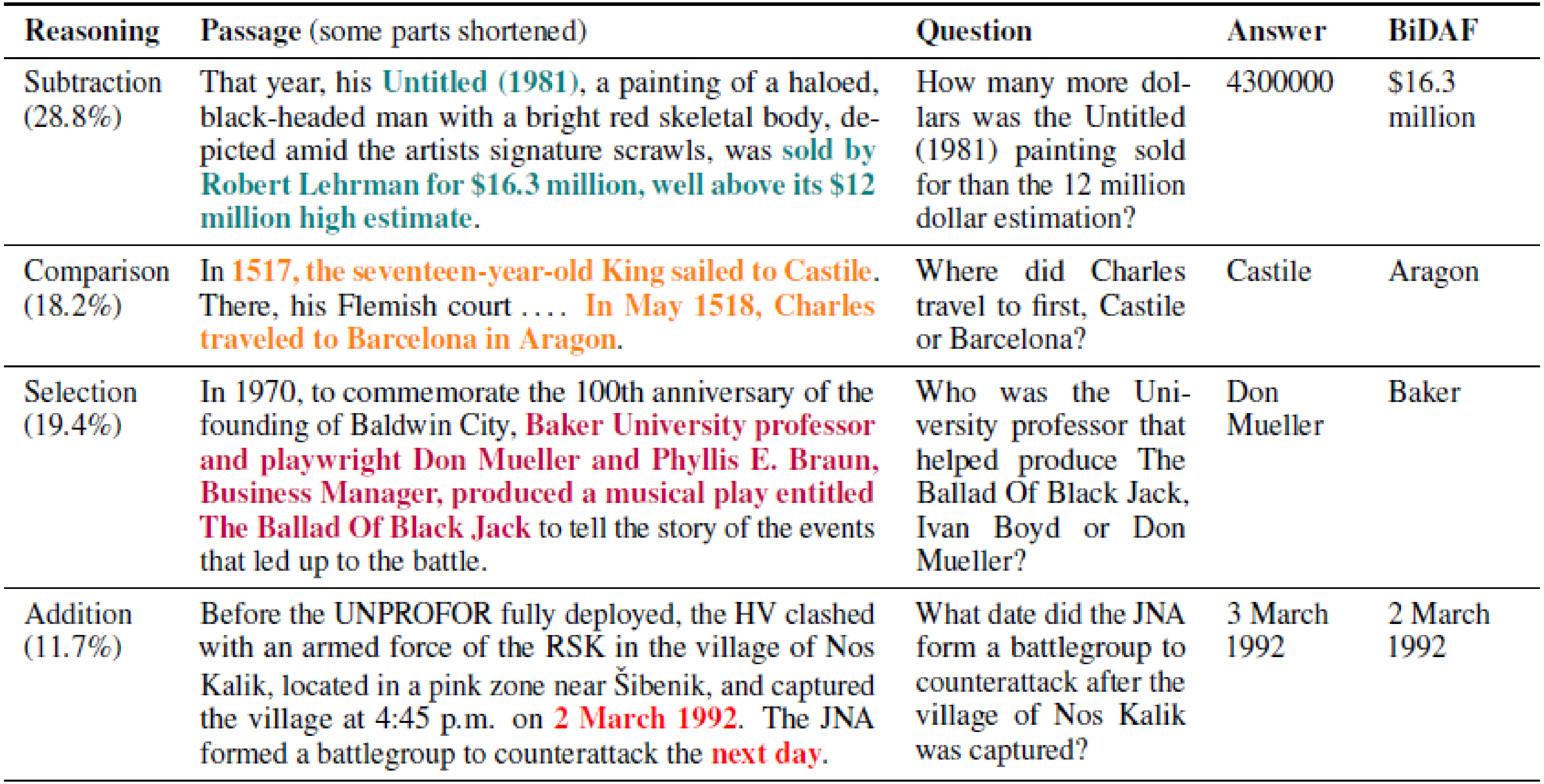
Introduction

- **DROP**
  - A system is given a paragraph and a question
  - Must perform some kind of Discrete Reasoning Over the text in the Paragraph to obtain the correct answer
  - Require discrete reasoning (such as addition, sorting, or counting; see Table 1)
    - These questions are inspired by the complex, compositional questions commonly found in the semantic parsing literature
  - Paragraph understanding
    - Complex questions allow us to test a system's understanding of the paragraph's semantics
  - Combine distributed representations with symbolic, discrete reasoning
    - System must be able to find multiple occurrences of an event described in a question (presumably using some kind of soft matching), extract arguments from the events
    - Perform a numerical operation such as a sort

Part 1.

Introduction

- **Dataset Construction**

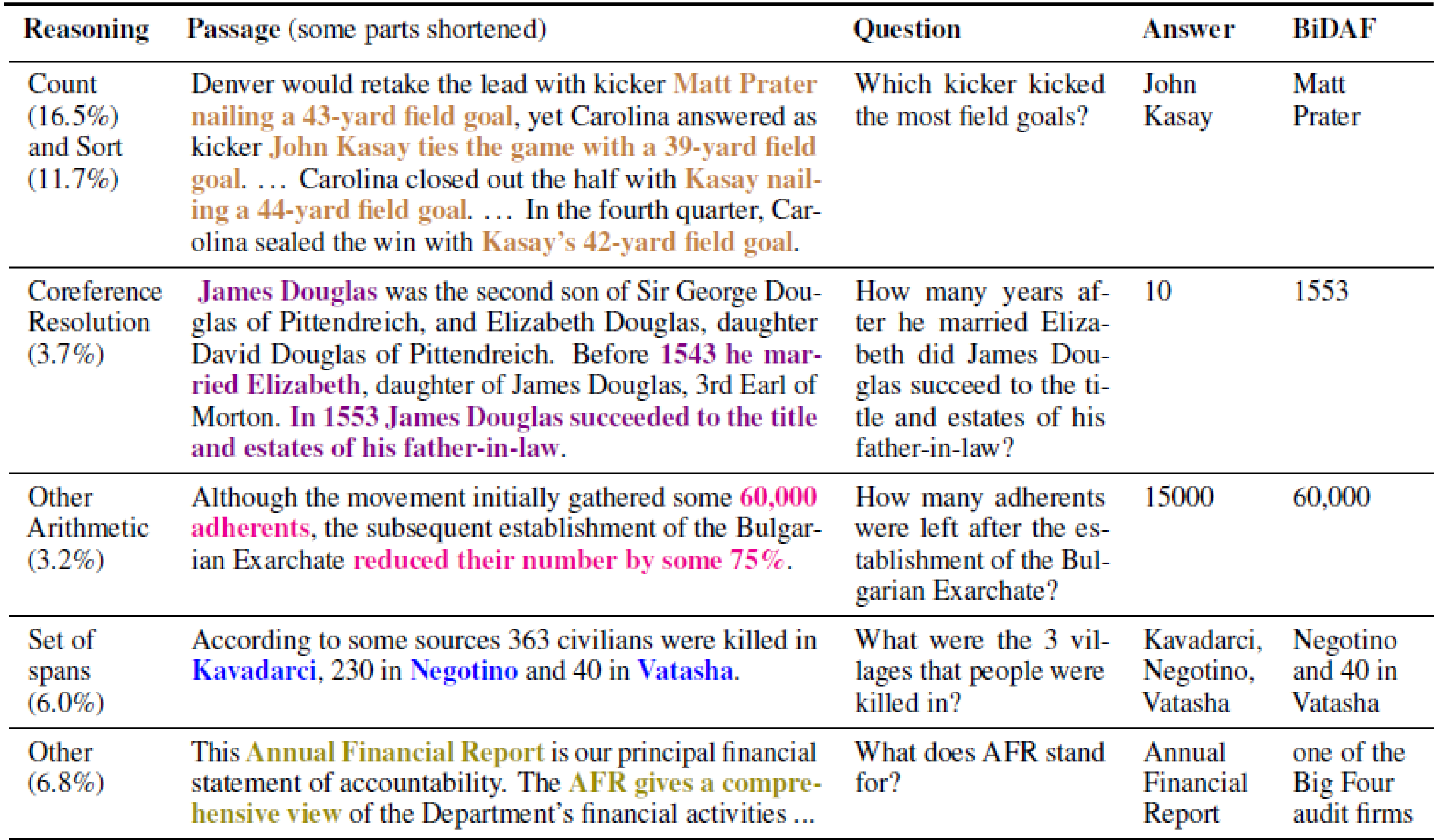


Reasoning	**Passage** (some parts shortened)	Question	Answer	BiDAF
Subtraction (28.8%)	That year, his Untitled (1981), a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was sold by Robert Lehrman for $16.3 million, well above its $12 million high estimate.	How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?	4300000	$16.3 million
Comparison (18.2%)	In 1517, the seventeen-year-old King sailed to Castile. There, his Flemish court .... In May 1518, Charles traveled to Barcelona in Aragon.	Where did Charles travel to first, Castile or Barcelona?	Castile	Aragon
Selection (19.4%)	In 1970, to commemorate the 100th anniversary of the founding of Baldwin City, Baker University professor and playwright Don Mueller and Phyllis E. Braun, Business Manager, produced a musical play entitled The Ballad Of Black Jack to tell the story of the events that led up to the battle.	Who was the University professor that helped produce The Ballad Of Black Jack, Ivan Boyd or Don Mueller?	Don Mueller	Baker
Addition (11.7%)	Before the UNPROFOR fully deployed, the HV clashed with an armed force of the RSK in the village of Nos Kalik, located in a pink zone near Šibenik, and captured the village at 4:45 p.m. on 2 March 1992. The JNA formed a battlegroup to counterattack the next day.	What date did the JNA form a battlegroup to counterattack after the village of Nos Kalik was captured?	3 March 1992	2 March 1992

Part 1.

Introduction

- **Dataset Construction**



Reasoning	**Passage** (some parts shortened)	**Question**	**Answer**	**BiDAF**
Count (16.5%) and Sort (11.7%)	Denver would retake the lead with kicker Matt Prater nailing a 43-yard field goal, yet Carolina answered as kicker John Kasay ties the game with a 39-yard field goal. … Carolina closed out the half with Kasay nailing a 44-yard field goal. … In the fourth quarter, Carolina sealed the win with Kasay's 42-yard field goal.	Which kicker kicked the most field goals?	John Kasay	Matt Prater
Coreference Resolution (3.7%)	James Douglas was the second son of Sir George Douglas of Pittendreich, and Elizabeth Douglas, daughter David Douglas of Pittendreich. Before 1543 he married Elizabeth, daughter of James Douglas, 3rd Earl of Morton. In 1553 James Douglas succeeded to the title and estates of his father-in-law.	How many years after he married Elizabeth did James Douglas succeed to the title and estates of his father-in-law?	10	1553
Other Arithmetic (3.2%)	Although the movement initially gathered some 60,000 adherents, the subsequent establishment of the Bulgarian Exarchate reduced their number by some 75%.	How many adherents were left after the establishment of the Bulgarian Exarchate?	15000	60,000
Set of spans (6.0%)	According to some sources 363 civilians were killed in Kavadarci, 230 in Negotino and 40 in Vatasha.	What were the 3 villages that people were killed in?	Kavadarci, Negotino, Vatasha	Negotino and 40 in Vatasha
Other (6.8%)	This Annual Financial Report is our principal financial statement of accountability. The AFR gives a comprehensive view of the Department's financial activities …	What does AFR stand for?	Annual Financial Report	one of the Big Four audit firms

Method

- **NAQANet**
  - DROP
    - Collect passages from Wikipedia that are easy to ask hard questions about
    - None of the baselines we described in Section 5 can do this
  - NAQANet: A numerically-aware QANet model
    - Allow the state-of-the-art reading comprehension system to produce three new answer types
    - (1) spans from the question; (2) counts; (3) addition or subtraction over numbers
  - Predict numbers
    - Predict whether the answer is a count or an arithmetic expression
    - Predict the specific numbers involved in the expression
  - The neural model producing a partially executed logical form
    - Combine neural methods and symbolic reasoning
    - Model is trained by marginalizing over all execution paths that lead to the correct answer

Method

- **Model Description**
  - Follow the typical architecture of previous reading comprehension models
    - It is composed of embedding, encoding, passage-question attention, and output layers
  - Use the original QANet architecture for everything up to the output layer
    - A question representation $\mathbf{Q} \in \mathbb{R}^{m \times d}$
    - A projected questionaware passage representation $\bar{\mathbf{P}} \in \mathbb{R}^{n \times d}$
  - Four different output layers
  - Four different kinds of answers the model can produce

Method

- **Passage span**
  - Predict an answer in the passage
    - Apply three repetitions of the QANet encoder to the passage representation $\bar{\mathbf{P}}$
    - Get their outputs as M_0, M_1
  - The probabilities of the starting and ending positions from the passage
    - FFN is a two-layer feed-forward network with the RELU activation

$$\mathbf{p}^{\text{p\_start}} = \text{softmax}(\text{FFN}([\mathbf{M}_0; \mathbf{M}_1])$$

$$\mathbf{p}^{\text{p\_end}} = \text{softmax}(\text{FFN}([\mathbf{M}_0; \mathbf{M}_2])$$

Method

- **Question span**
  - Some questions in DROP have their answer in the question instead of the passage
  - Predict an answer from the question,
    - Computes a vector h^P that represents the information it finds in the passage

$$\boldsymbol{\alpha}^P = \mathrm{softmax}(\mathbf{W}^P\bar{\mathbf{P}})$$

$$\mathbf{h}^P = \boldsymbol{\alpha}^P\bar{\mathbf{P}}$$

  - The probabilities of the starting and ending positions from the question
    - The outer product with the identity $(\mathbf{e}^{|Q|} \otimes \cdot)$ simply repeats h^P for each question word

$$\mathbf{p}^{\text{q\_start}} = \mathrm{softmax}(\mathrm{FFN}([\mathbf{Q}; \mathbf{e}^{|Q|} \otimes \mathbf{h}^P])$$

$$\mathbf{p}^{\text{q\_end}} = \mathrm{softmax}(\mathrm{FFN}([\mathbf{Q}; \mathbf{e}^{|Q|} \otimes \mathbf{h}^P])$$

Method

- **Count**
  - Model the capability of counting as a multi-class classification problem
  - Consider ten numbers (0–9) in this preliminary model and the probabilities of choosing these numbers is computed based on the passage vector h^P

$$\mathbf{p}^{\text{count}} = \text{softmax}(\text{FFN}(\mathbf{h}^P))$$

Part 2.

Method

- **Arithmetic expression**
  - Get an arithmetic expression composed of signed numbers, which can be evaluated to give the final answer
    - Extract all the numbers from the passage
    - Learn to assign a plus, minus or zero for each number
  - Apply another QANet encoder to M_2 and get a new passage representation M_3
  - Select an index over the concatenation of M_0 and M_3, to get a representation for each number in this passage
  - The i^{th} number can be represented as h_i^N and the probabilities of this number being assigned a plus, minus or zero are

$$\mathbf{p}_i^{\text{sign}} = \text{softmax}(\text{FFN}(\mathbf{h}_i^N))$$

Part 2.

Method

- **Answer type prediction**
  - Use a categorical variable to decide between the above four answer types, with probabilities computed as:
    - Computes a vector h^P that represents the information it finds in the passage
    - h^Q is computed over Q, in a similar way as we did for h^P

$$\mathbf{p}^{\text{type}} = \text{softmax}(\text{FFN}([\mathbf{h}^P, \mathbf{h}^Q]))$$

  - Test time
    - Determine this answer type greedily
    - Get the best answer from the selected type Computes

Method

- **Weakly-Supervised Training**
  - DROP
    - Contain only the answer string
    - Not which of the above answer types is used to arrive at the answer
  - To train our model
    - Adopt the weakly supervised training method widely used in the semantic parsing literature (Berant et al., 2013a)
  - Find all executions that evaluate to the correct answer
    - Match passage spans and question spans
    - Correct count numbers
    - Sign assignments for numbers
  - Training objective
    - Maximize the marginal likelihood of these executions

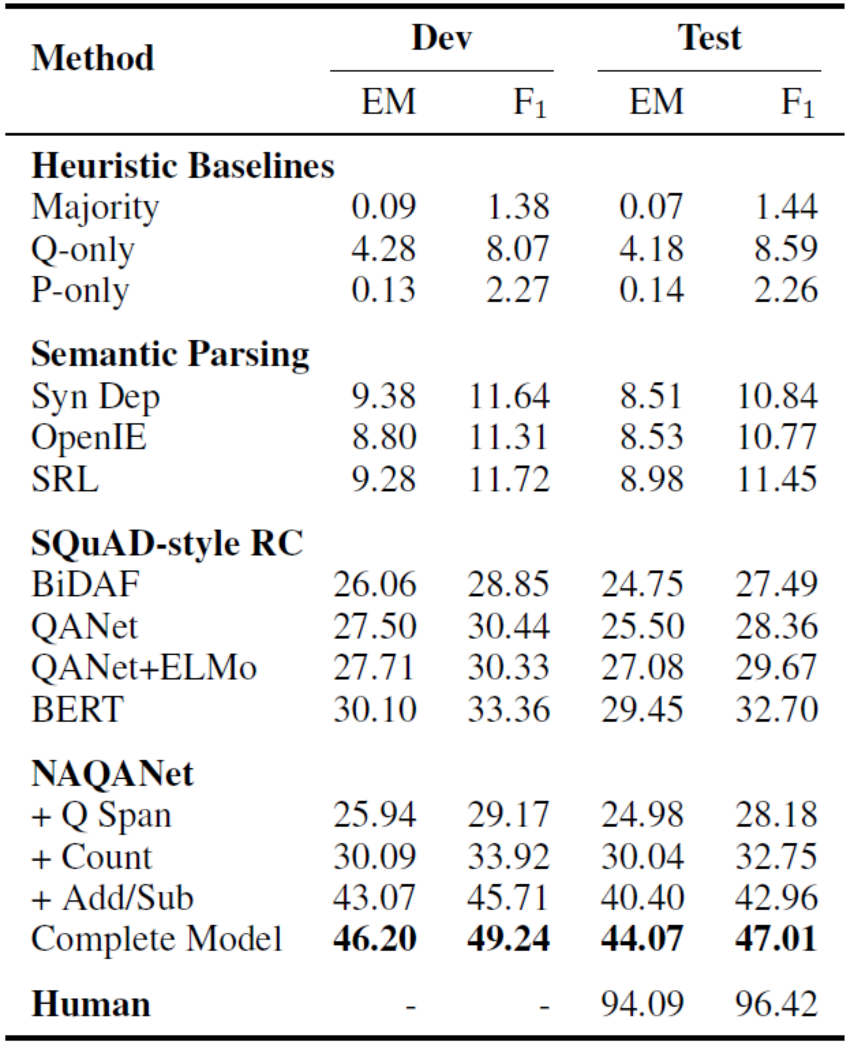
Experiment

- **Results and Discussion**
  - Llemma family (Azerbayev et al., 2023, finetuned for math and reasoning)
  - The MEDITRON family (Chen et al., 2023, finetuned for biomedicine)
  - The 7B and the 70B model from the same LLAMA2 family (Touvron et al., 2023) as the base and assistant model

- **Datasets**
  - Instruction following
    - Full Tülu v2 mix data (Wang et al., 2023) for
  - Reasoning and math problem solving
    - GSM8k (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021)
  - Medical question answering
    - The BioASQ (Tsatsaronis et al., 2015) (4.7k samples)

Experiment

- **The performance of all tested models on the DROP dataset**



Method	Dev		Test	
	EM	F_1	EM	F_1
Heuristic Baselines				
Majority	0.09	1.38	0.07	1.44
Q-only	4.28	8.07	4.18	8.59
P-only	0.13	2.27	0.14	2.26
Semantic Parsing				
Syn Dep	9.38	11.64	8.51	10.84
OpenIE	8.80	11.31	8.53	10.77
SRL	9.28	11.72	8.98	11.45
SQuAD-style RC				
BiDAF	26.06	28.85	24.75	27.49
QANet	27.50	30.44	25.50	28.36
QANet+ELMo	27.71	30.33	27.08	29.67
BERT	30.10	33.36	29.45	32.70
NAQANet				
+ Q Span	25.94	29.17	24.98	28.18
+ Count	30.09	33.92	30.04	32.75
+ Add/Sub	43.07	45.71	40.40	42.96
Complete Model	**46.20**	**49.24**	**44.07**	**47.01**
Human	-	-	94.09	96.42

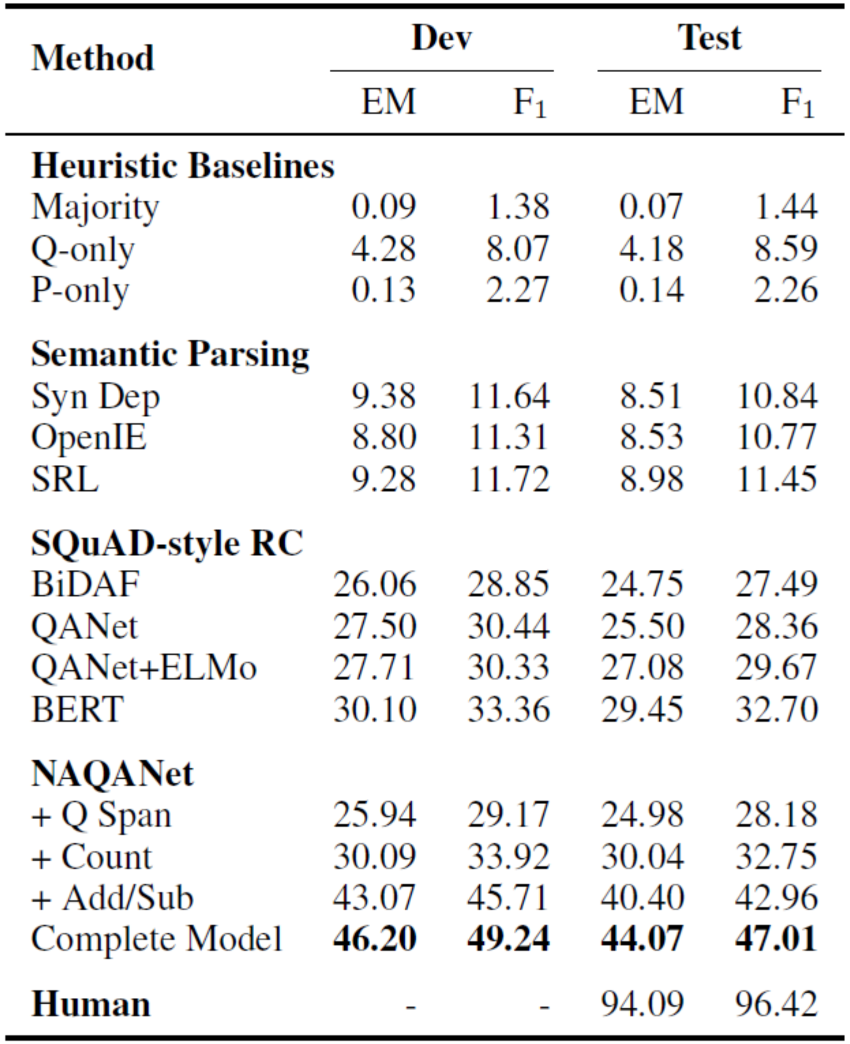
- Drop: a challenging reading comprehension dataset
  - All models perform significantly worse than on other prominent reading comprehension datasets
  - Human performance remains at similar high levels

- The best performance is obtained by our NAQANet model

Part 3.

Experiment

- **Difficulties of building semantic parsers**

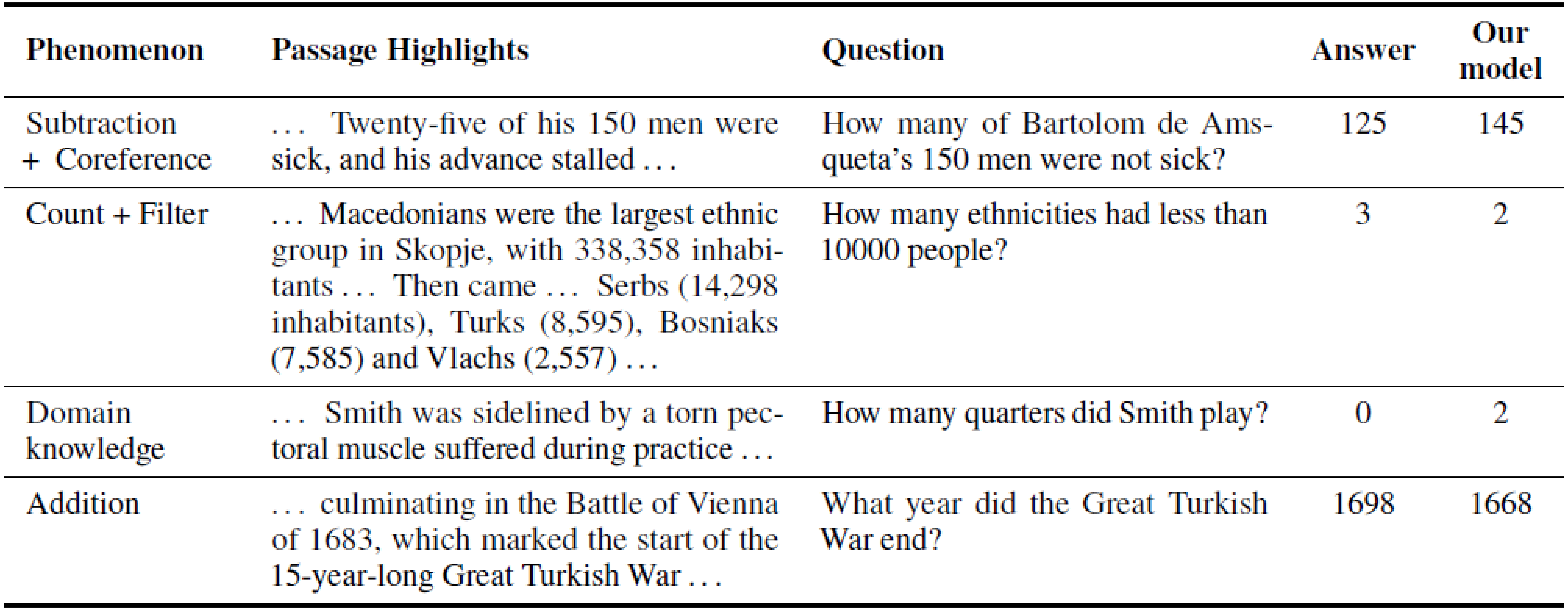


Method	Dev		Test	
	EM	F_1	EM	F_1
Heuristic Baselines				
Majority	0.09	1.38	0.07	1.44
Q-only	4.28	8.07	4.18	8.59
P-only	0.13	2.27	0.14	2.26
Semantic Parsing				
Syn Dep	9.38	11.64	8.51	10.84
OpenIE	8.80	11.31	8.53	10.77
SRL	9.28	11.72	8.98	11.45
SQuAD-style RC				
BiDAF	26.06	28.85	24.75	27.49
QANet	27.50	30.44	25.50	28.36
QANet+ELMo	27.71	30.33	27.08	29.67
BERT	30.10	33.36	29.45	32.70
NAQANet				
+ Q Span	25.94	29.17	24.98	28.18
+ Count	30.09	33.92	30.04	32.75
+ Add/Sub	43.07	45.71	40.40	42.96
Complete Model	**46.20**	**49.24**	**44.07**	**47.01**
Human	-	-	94.09	96.42

- All the semantic parsing baselines perform quite poorly
  - Our pipeline of extracting tabular information from paragraphs, followed by the denotation-driven logical form search, can yield logical forms only for a subset of the training data
- High quality information extraction is a strong prerequisite for building semantic parsers
  - This is a weakly supervised semantic parsing problem also makes training hard
- The spuriousness of logical forms used for training
  - The logical form evaluates to the correct denotation but does not actually reflect the semantics of the question
  - This makes it hard for the model trained on these spurious logical forms to generalize to unseen data

Part 3. Experiment

- **Error Analysis**
  - Analysis on a random sample of 100 erroneous NAQANet predictions
  - Arithmetic operations (51%), counting (30%), domain knowledge and common sense (23%), co-reference (6%), or a combination of different types of reasoning (40%)



Phenomenon	Passage Highlights	Question	Answer	Our model
Subtraction + Coreference	… Twenty-five of his 150 men were sick, and his advance stalled …	How many of Bartolom de Amsqueta's 150 men were not sick?	125	145
Count + Filter	… Macedonians were the largest ethnic group in Skopje, with 338,358 inhabitants … Then came … Serbs (14,298 inhabitants), Turks (8,595), Bosniaks (7,585) and Vlachs (2,557) …	How many ethnicities had less than 10000 people?	3	2
Domain knowledge	… Smith was sidelined by a torn pectoral muscle suffered during practice …	How many quarters did Smith play?	0	2
Addition	… culminating in the Battle of Vienna of 1683, which marked the start of the 15-year-long Great Turkish War …	What year did the Great Turkish War end?	1698	1668

Part 4.

Conclusion

- **DROP: Discrete Reasoning Over Paragraphs**
  - Substantially more challenging than existing datasets
  - The best baseline achieving only 32.7% F1, while humans achieve 96%

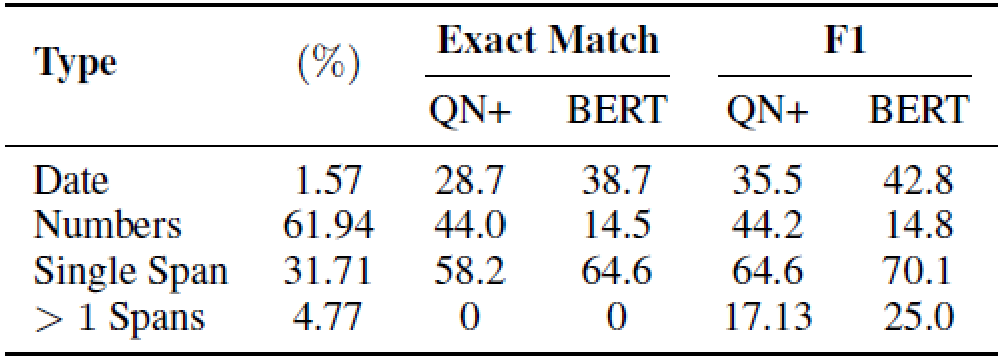
  - Hope
    - Spur research into more comprehensive analysis of paragraphs
    - Combine distributed representations with symbolic reasoning

  - We have Additionally presented initial work in this direction
    - Augment QANet with limited numerical reasoning capability
    - Achieve 47% F1 on DROP.

Part 5.

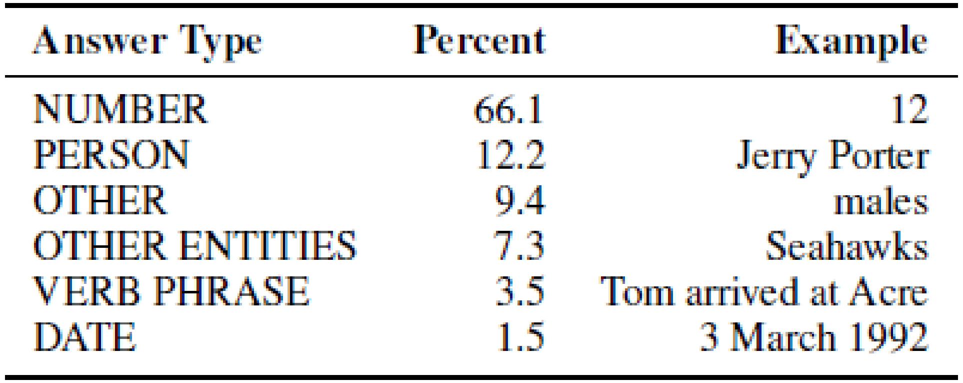
Appendix

• DROP Data Analysis



Type	(%)	Exact Match		F1	
		QN+	BERT	QN+	BERT
Date	1.57	28.7	38.7	35.5	42.8
Numbers	61.94	44.0	14.5	44.2	14.8
Single Span	31.71	58.2	64.6	64.6	70.1
> 1 Spans	4.77	0	0	17.13	25.0

Dataset statistics across the different splits



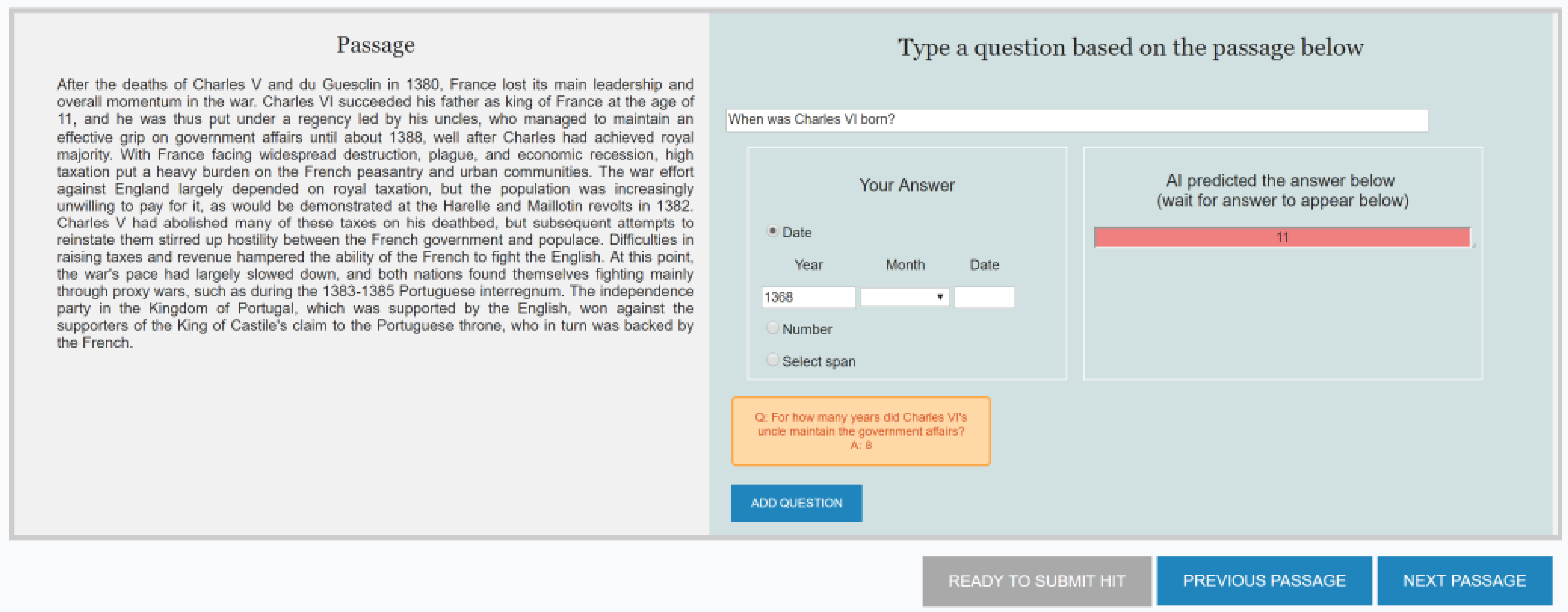
Answer Type	Percent	Example
NUMBER	66.1	12
PERSON	12.2	Jerry Porter
OTHER	9.4	males
OTHER ENTITIES	7.3	Seahawks
VERB PHRASE	3.5	Tom arrived at Acre
DATE	1.5	3 March 1992

Distribution of answer types in training set, according to an automatic named entity recognition

Part 5.

Appendix

- **Question Answering HIT sample above with passage**



Passage

After the deaths of Charles V and du Guesclin in 1380, France lost its main leadership and overall momentum in the war. Charles VI succeeded his father as king of France at the age of 11, and he was thus put under a regency led by his uncles, who managed to maintain an effective grip on government affairs until about 1388, well after Charles had achieved royal majority. With France facing widespread destruction, plague, and economic recession, high taxation put a heavy burden on the French peasantry and urban communities. The war effort against England largely depended on royal taxation, but the population was increasingly unwilling to pay for it, as would be demonstrated at the Harelle and Maillotin revolts in 1382. Charles V had abolished many of these taxes on his deathbed, but subsequent attempts to reinstate them stirred up hostility between the French government and populace. Difficulties in raising taxes and revenue hampered the ability of the French to fight the English. At this point, the war's pace had largely slowed down, and both nations found themselves fighting mainly through proxy wars, such as during the 1383-1385 Portuguese interregnum. The independence party in the Kingdom of Portugal, which was supported by the English, won against the supporters of the King of Castile's claim to the Portuguese throne, who in turn was backed by the French.

Type a question based on the passage below

When was Charles VI born?

Your Answer

● Date

Year	Month	Date
1368	▾	

○ Number

○ Select span

AI predicted the answer below
(wait for answer to appear below)

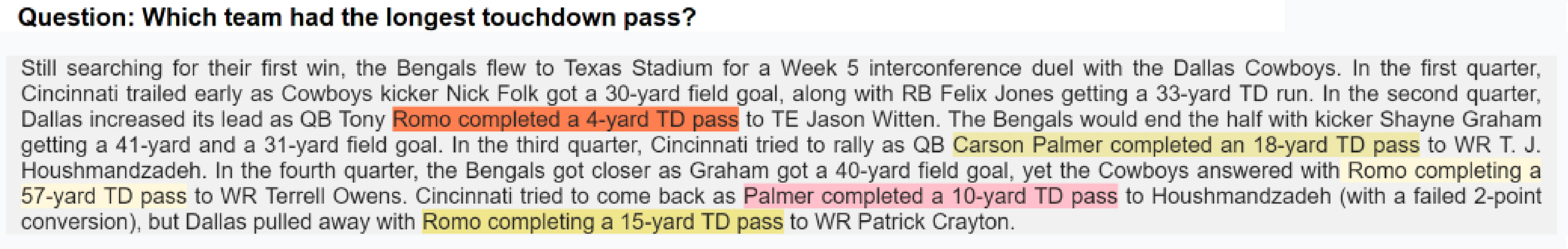
11

Q: For how many years did Charles VI's uncle maintain the government affairs?
A: 8

Part 5.

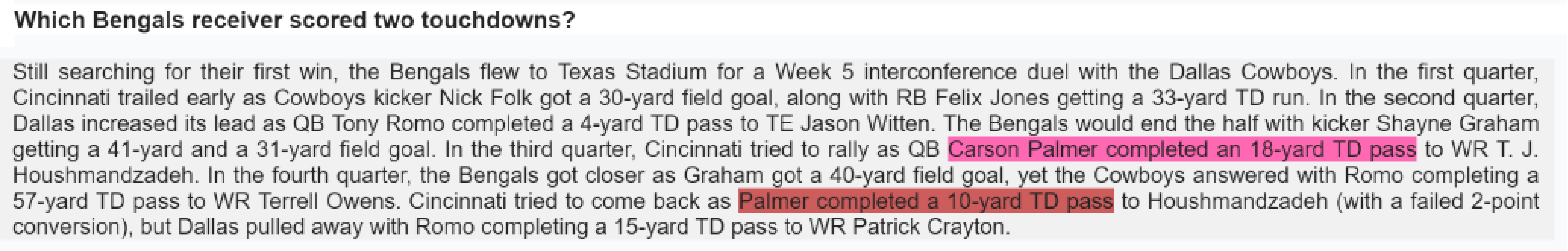
Appendix

- **Question Answering HIT sample above with passage**



Question: Which team had the longest touchdown pass?

Still searching for their first win, the Bengals flew to Texas Stadium for a Week 5 interconference duel with the Dallas Cowboys. In the first quarter, Cincinnati trailed early as Cowboys kicker Nick Folk got a 30-yard field goal, along with RB Felix Jones getting a 33-yard TD run. In the second quarter, Dallas increased its lead as QB Tony Romo completed a 4-yard TD pass to TE Jason Witten. The Bengals would end the half with kicker Shayne Graham getting a 41-yard and a 31-yard field goal. In the third quarter, Cincinnati tried to rally as QB Carson Palmer completed an 18-yard TD pass to WR T. J. Houshmandzadeh. In the fourth quarter, the Bengals got closer as Graham got a 40-yard field goal, yet the Cowboys answered with Romo completing a 57-yard TD pass to WR Terrell Owens. Cincinnati tried to come back as Palmer completed a 10-yard TD pass to Houshmandzadeh (with a failed 2-point conversion), but Dallas pulled away with Romo completing a 15-yard TD pass to WR Patrick Crayton.



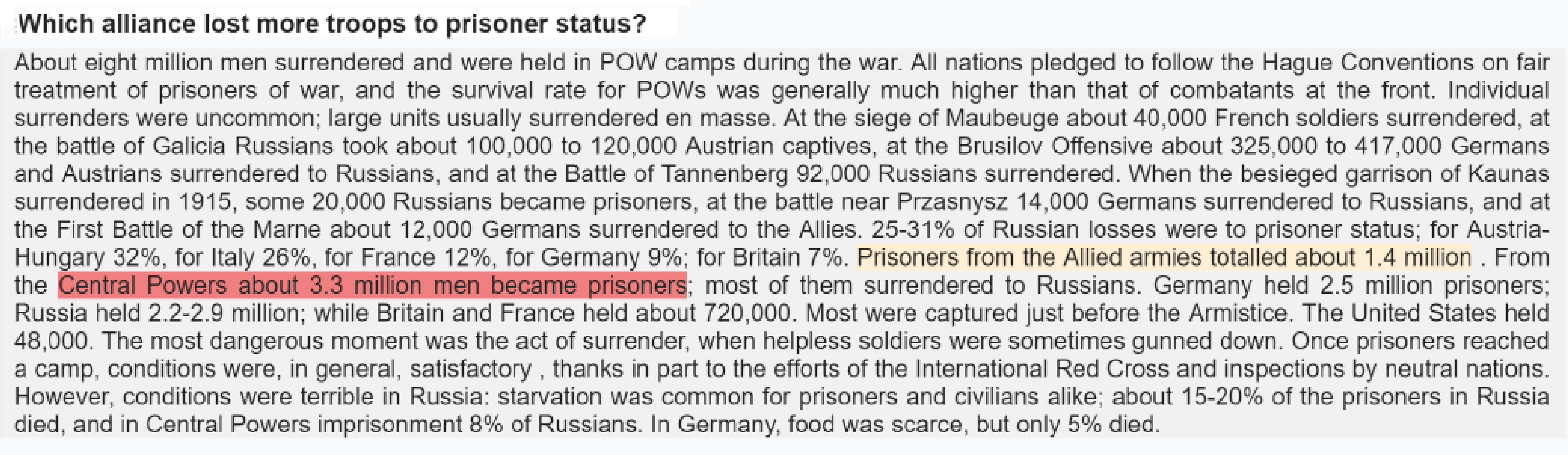
Which Bengals receiver scored two touchdowns?

Still searching for their first win, the Bengals flew to Texas Stadium for a Week 5 interconference duel with the Dallas Cowboys. In the first quarter, Cincinnati trailed early as Cowboys kicker Nick Folk got a 30-yard field goal, along with RB Felix Jones getting a 33-yard TD run. In the second quarter, Dallas increased its lead as QB Tony Romo completed a 4-yard TD pass to TE Jason Witten. The Bengals would end the half with kicker Shayne Graham getting a 41-yard and a 31-yard field goal. In the third quarter, Cincinnati tried to rally as QB Carson Palmer completed an 18-yard TD pass to WR T. J. Houshmandzadeh. In the fourth quarter, the Bengals got closer as Graham got a 40-yard field goal, yet the Cowboys answered with Romo completing a 57-yard TD pass to WR Terrell Owens. Cincinnati tried to come back as Palmer completed a 10-yard TD pass to Houshmandzadeh (with a failed 2-point conversion), but Dallas pulled away with Romo completing a 15-yard TD pass to WR Patrick Crayton.

Part 5.

Appendix

- **Question Answering HIT sample above with passage**



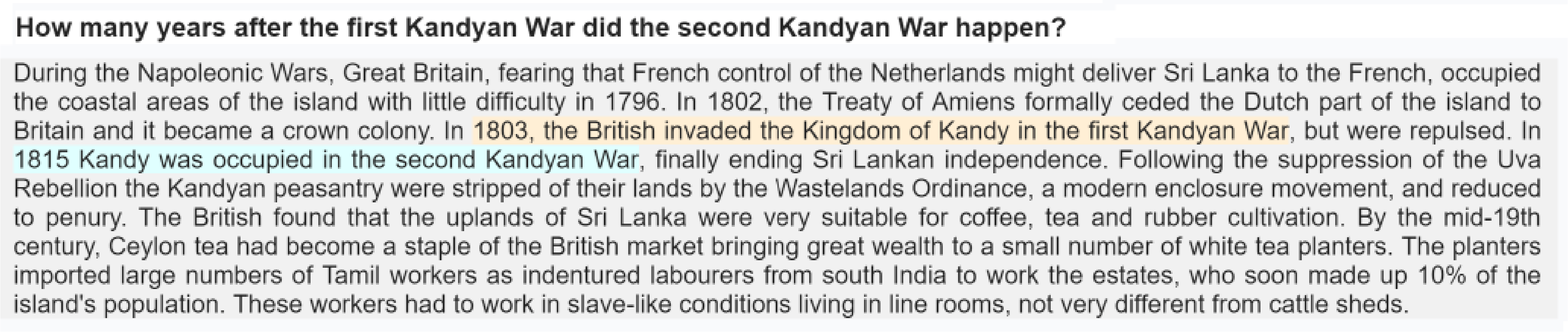
Which alliance lost more troops to prisoner status?

About eight million men surrendered and were held in POW camps during the war. All nations pledged to follow the Hague Conventions on fair treatment of prisoners of war, and the survival rate for POWs was generally much higher than that of combatants at the front. Individual surrenders were uncommon; large units usually surrendered en masse. At the siege of Maubeuge about 40,000 French soldiers surrendered, at the battle of Galicia Russians took about 100,000 to 120,000 Austrian captives, at the Brusilov Offensive about 325,000 to 417,000 Germans and Austrians surrendered to Russians, and at the Battle of Tannenberg 92,000 Russians surrendered. When the besieged garrison of Kaunas surrendered in 1915, some 20,000 Russians became prisoners, at the battle near Przasnysz 14,000 Germans surrendered to Russians, and at the First Battle of the Marne about 12,000 Germans surrendered to the Allies. 25-31% of Russian losses were to prisoner status; for Austria-Hungary 32%, for Italy 26%, for France 12%, for Germany 9%; for Britain 7%. Prisoners from the Allied armies totalled about 1.4 million . From the Central Powers about 3.3 million men became prisoners; most of them surrendered to Russians. Germany held 2.5 million prisoners; Russia held 2.2-2.9 million; while Britain and France held about 720,000. Most were captured just before the Armistice. The United States held 48,000. The most dangerous moment was the act of surrender, when helpless soldiers were sometimes gunned down. Once prisoners reached a camp, conditions were, in general, satisfactory , thanks in part to the efforts of the International Red Cross and inspections by neutral nations. However, conditions were terrible in Russia: starvation was common for prisoners and civilians alike; about 15-20% of the prisoners in Russia died, and in Central Powers imprisonment 8% of Russians. In Germany, food was scarce, but only 5% died.

Part 5.

Appendix

- **Question Answering HIT sample above with passage**



How many years after the first Kandyan War did the second Kandyan War happen?

During the Napoleonic Wars, Great Britain, fearing that French control of the Netherlands might deliver Sri Lanka to the French, occupied the coastal areas of the island with little difficulty in 1796. In 1802, the Treaty of Amiens formally ceded the Dutch part of the island to Britain and it became a crown colony. In 1803, the British invaded the Kingdom of Kandy in the first Kandyan War, but were repulsed. In 1815 Kandy was occupied in the second Kandyan War, finally ending Sri Lankan independence. Following the suppression of the Uva Rebellion the Kandyan peasantry were stripped of their lands by the Wastelands Ordinance, a modern enclosure movement, and reduced to penury. The British found that the uplands of Sri Lanka were very suitable for coffee, tea and rubber cultivation. By the mid-19th century, Ceylon tea had become a staple of the British market bringing great wealth to a small number of white tea planters. The planters imported large numbers of Tamil workers as indentured labourers from south India to work the estates, who soon made up 10% of the island's population. These workers had to work in slave-like conditions living in line rooms, not very different from cattle sheds.