Question answering

- IR-based QA
- Commonsense QA
- Knowledge-based QA
- Logical reasoning QA

Types of questions

1. Factoid questions:
   - What is the dress code for the Vatican?
   - Who is the President of the United States?
   - What are the dots in Hebrew called?
2. Commonsense questions:
   - What do all humans want to experience in their own home? (a) feel comfortable, (b) work hard, (c) fall in love, (d) lay eggs, (e) live forever
3. Opinion questions:
   - Can anyone recommend a good coffee shop near HSE campus?
4. Cloze-style questions

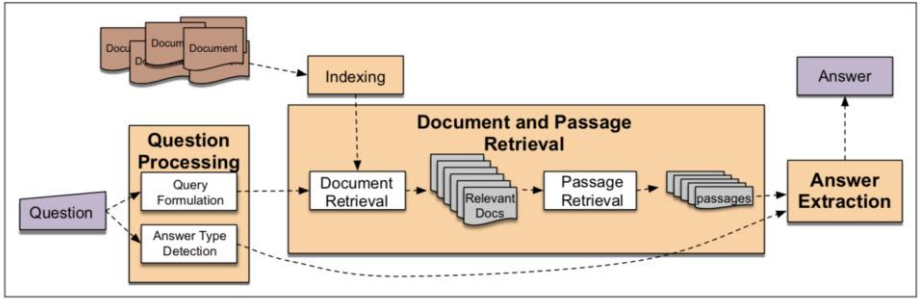
Types of questions

1. Types of answers
   - binary (yes / now)
   - find a span of text
   - multiple choice

Major paradigms for factoid question answering

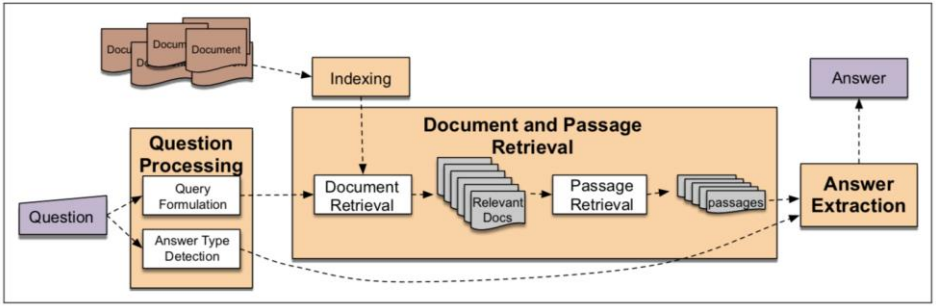
1. Information retrieval (IR)-based QA: find a span of text, which answers a question
2. Open-domain Question Answering (ODQA): answer questions about nearly anything
3. Knowledge (KB)-based QA: build a semantic representation of question are used to question knowledge bases
   When Bernardo Bertolucci died? $\rightarrow$ `death-year(Bernardo Bertolucci, ?x)`

IR-based QA



1. Question processing
   - answer type (PER, LOC, TIME)
   - focus
   - question type
2. Query formulation
   - question reformulation: remove *wh*-words, change word order
   - query expansion

IR-based QA



3. Document and passage retrieval
4. Answer extraction
   What are the dots in Hebrew called?
   In Hebrew orthography, **niqqud or nikkud**, *is a system of diacritical signs used to represent vowels or distinguish between alternative pronunciations of letters of the Hebrew alphabet.*

Datasets for IR-based QA

Passage: Tesla later approached Morgan to ask for more funds to build a more powerful transmitter. **When asked where all the money had gone, Tesla responded by saying that he was affected by the Panic of 1901**, which he (Morgan) had caused. Morgan was shocked by the reminder of his part in the stock market crash and by Tesla's breach of contract by asking for more funds. Tesla wrote another plea to Morgan, but it was also fruitless. Morgan still owed Tesla money on the original agreement, and Tesla had been facing foreclosure even before construction of the tower began.

Question: On what did Tesla blame for the loss of the initial money?

Answer: Panic of 1901

Figure: An example from the SQuAD dataset

1. Stanford Question Answering Dataset (SQuAD)
2. NewsQA
3. WikiQA
4. CuratedTREC
5. WebQuestions
6. WikiMovies
7. Russian: SberQUAD

SQuAD2.0

100,000 questions in SQuAD1.1 and over 50,000 unaswerable questions in SQuAS2.0

1. Project Nayuki's Wikipedia's internal PageRanks to obtain the top 10000 articles of English Wikipedia, from which we sampled 536 articles uniformly at random
2. Articles splitted in individual paragraphes
3. Crowsourcing: ask and answer up to 5 questions on the content of that paragraph
4. Crowdworkers were encouraged to ask questions in their own words, without copying word phrases from the paragraph
5. Analysis: the (i) diversity of answer types, (ii) the difficulty of questions in terms of type of reasoning required to answer them, and (iii) the degree of syntactic divergence between the question and answer sentences.

`https://rajpurkar.github.io/SQuAD-explorer/`

RACE

Passage:
In a small village in England about 150 years ago, a mail coach was standing on the street. It didn't come to that village often. People had to pay a lot to get a letter. The person who sent the letter didn't have to pay the postage, while the receiver had to.
"Here's a letter for Miss Alice Brown," said the mailman.
" I'm Alice Brown," a girl of about 18 said in a low voice.
Alice looked at the envelope for a minute, and then handed it back to the mailman.
"I'm sorry I can't take it, I don't have enough money to pay it", she said.
A gentleman standing around were very sorry for her. Then he came up and paid the postage for her.
When the gentleman gave the letter to her, she said with a smile, " Thank you very much, This letter is from Tom. I'm going to marry him. He went to London to look for work. I've waited a long time for this letter, but now I don't need it, there is nothing in it."
"Really? How do you know that?" the gentleman said in surprise.
"He told me that he would put some signs on the envelope. Look, sir, this cross in the corner means that he is well and this circle means he has found work. That's good news."
The gentleman was Sir Rowland Hill. He didn't forgot Alice and her letter.
"The postage to be paid by the receiver has to be changed," he said to himself and had a good plan.
"The postage has to be much lower, what about a penny? And the person who sends the letter pays the postage. He has to buy a stamp and put it on the envelope." he said . The government accepted his plan. Then the first stamp was put out in 1840. It was called the "Penny Black". It had a picture of the Queen on it.

Questions:

1): The first postage stamp was made ..
A. in England B. in America C. by Alice D. in 1910

2): The girl handed the letter back to the mailman because . .
A. she didn't know whose letter it was
B. she had no money to pay the postage
C. she received the letter but she didn't want to open it
D. she had already known what was written in the letter

3): We can know from Alice's words that . .
A. Tom had told her what the signs meant before leaving
B. Alice was clever and could guess the meaning of the signs
C. Alice had put the signs on the envelope herself
D. Tom had put the signs as Alice had told him to

4): The idea of using stamps was thought of by . .
A. the government
B. Sir Rowland Hill
C. Alice Brown
D. Tom

5): From the passage we know the high postage made . .
A. people never send each other letters
B. lovers almost lose every touch with each other
C. people try their best to avoid paying it
D. receivers refuse to pay the coming letters

Answer: ADABC

Figure: An example from RACE dataset

RACE consists of near 28k passages and near 100k questions generated by human experts (English instructors), and covers a variety of topics which are carefully designed for evaluating the students' ability in understanding and reasoning.

RACE

Dataset	RACE-M	RACE-H	RACE	CNN	SQUAD	NEWSQA
Word Matching	29.4%	11.3%	15.8%	13.0%†	39.8%*	32.7%*
Paraphrasing	14.8%	20.6%	19.2%	41.0%†	34.3%*	27.0%*
Single-Sentence Reasoning	31.3%	34.1%	33.4%	19.0%†	8.6%*	13.2%*
Multi-Sentence Reasoning	22.6%	26.9%	25.8%	2.0%†	11.9%*	20.7%*
Ambiguous/Insufficient	1.8%	7.1%	5.8%	25.0%†	5.4%*	6.4%*

Figure: Statistic information about Reasoning type in different datasets

RACE includes five classes of questions: word matching, paraphrasing, single-sentence reasoning, multi-sentence reasoning, insufficient or ambiguous questions.
`http://www.cs.cmu.edu/~glai1/data/race/`

MS Marco

Field	Description
Query	A question query issued to Bing.
Passages	Top 10 passages from Web documents as retrieved by Bing. The passages are presented in ranked order to human editors. The passage that the editor uses to compose the answer is annotated as is_selected: 1.
Document URLs	URLs of the top ranked documents for the question from Bing. The passages are extracted from these documents.
Answer(s)	Answers composed by human editors for the question, automatically extracted passages and their corresponding documents.
Well Formed Answer(s)	Well-formed answer rewritten by human editors, and the original answer.
Segment	QA classification. E.g., tallest mountain in south america belongs to the ENTITY segment because the answer is an entity (Aconcagua).

Figure: The final dataset format for MS MARCO

Three tasks:

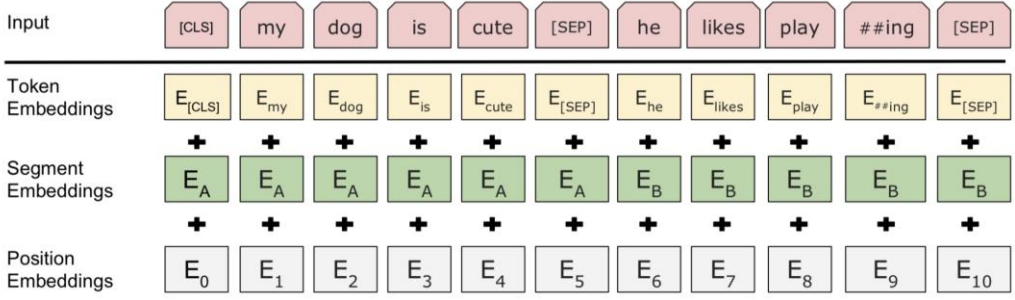
1. first predict whether a question can be answered, if so, generate the correct answer
2. the generated answer should be well-formed
3. the passage re-ranking

`http://www.msmarco.org`

BERT

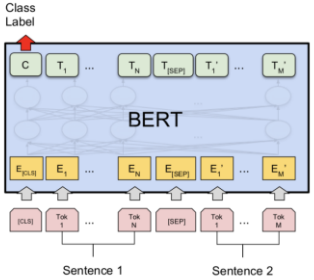
Bidirectional Encoder Representations from Transformers

- L – number of Transformer blocks, H – hidden size, A – the number of self-attention heads
- BERT$_{BASE}$: L=12, H=768, A =12, Total Parameters=110M
- Embeddings: WordPiece + position + segment
- **Two tasks**: Masked LM, Next Sentence Prediction

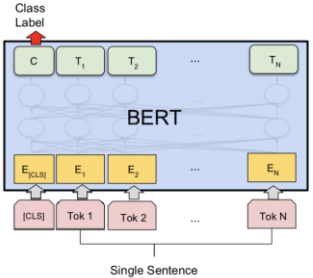


BERT

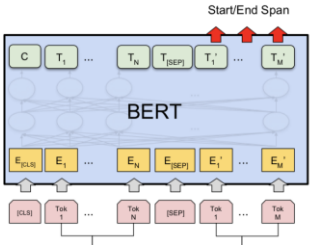
Bidirectional Encoder Representations from Transformers

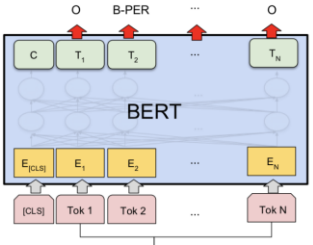


(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA





CommonsenseQA

Where would I not want a fox?
👍 hen house, 👎 england, 👎 mountains,
👎 english hunt, 👎 california

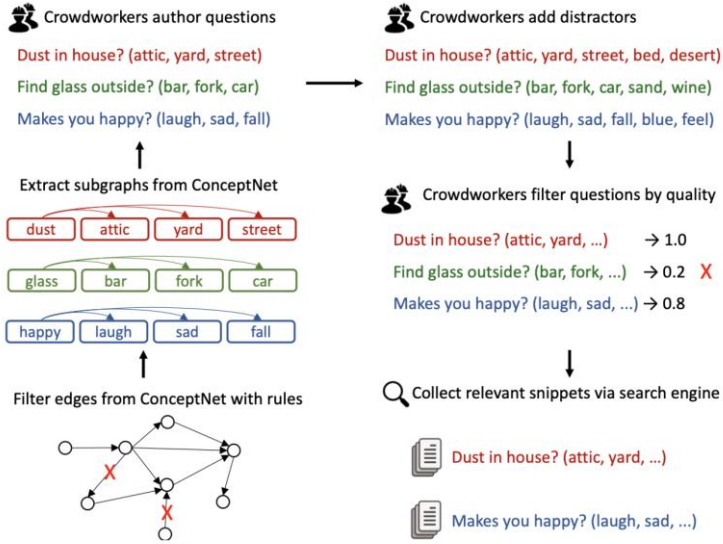
Why do people read gossip magazines?
👍 entertained, 👎 get information, 👎 learn,
👎 improve know how, 👎 lawyer told to

What do all humans want to experience in their own home?
👍 feel comfortable, 👎 work hard, 👎 fall in love,
👎 lay eggs, 👎 live forever

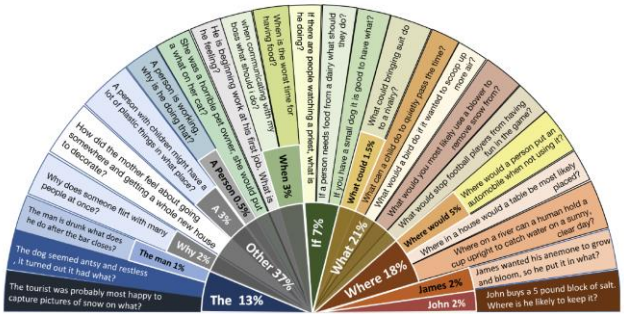
12,247 multiple choice questions that require common sense understanding

`https://www.tau-nlp.org/commonsenseqa`

CommonsenseQA



CommonsenseQA



Baseline: apply BERT

1. Linearize each question: [CLS] If ... ? [SEP] bedroom [SEP]
2. Fine-tune the pre-trained weights from BERT model
3. The hidden representations over each [CLS] token are run through a softmax layer to create the predictions

SWAG

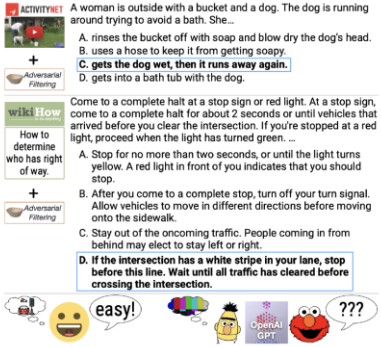
On stage, a woman takes a seat at the piano. She
a) sits on a bench as her sister plays with the doll.
b) smiles with someone as the music plays.
c) is in the crowd, watching the dancers.
d) nervously sets her fingers on the keys.

A girl is going across a set of monkey bars. She
a) jumps up across the monkey bars.
b) struggles onto the monkey bars to grab her head.
c) gets to the end and stands on a wooden plank.
d) jumps up and does a back flip.

The woman is now blow drying the dog. The dog
a) is placed in the kennel next to a woman's feet.
b) washes her face with the shampoo.
c) walks into frame and walks towards the dog.
d) tried to cut her face, so she is trying to do something very close to her face.

SWAG is a dataset for studying grounded commonsense inference. It consists of 113k multiple choice questions about grounded situations: each question comes from a video caption, with four answer choices about what might happen next in the scene. The correct answer is the (real) video caption for the next event in the video; the three incorrect answers are adversarially generated and human verified, so as to fool machines but not humans. `https://rowanzellers.com/swag/`

HellaSWAG



HellaSWAG is a dataset for studying grounded commonsense inference. It consists of 70k multiple choice questions about grounded situations: each question comes from one of two domains – activitynet or wikihow – with four answer choices about what might happen next in the scene. The correct answer is the (real) sentence for the next event; the three incorrect answers are adversarially generated and human verified, so as to fool machines but not humans. `https://rowanzellers.com/hellaswag/`

Knowledge-based QA

subject	predicate	object
Lyubov Polishchuk	death-date	28 November 2006

- When Lyubov Polishchuk died?
- Who died on 28 November 2006?

1. **Rule-based methods**: patterns that search for the question word and main verb
2. **OpenIE**: map between the words in question and canonical relations
3. **Knowledge base / knowledge graph**: match the words to concepts and relations in KB / KG

Knowledge representation

"leonardo da vinci"
String

Leonardo da Vinci
Recognized entity

Leonardo da Vinci
Recognized entity
Related entities

Leonardo da Vinci
Recognized entity
Related entities
Named Relationship

medium

Datasets

What American cartoonist is the creator of Andy Lippincott?	(andy_lippincott, character_created_by, garry_trudeau)
Which forest is Fires Creek in?	(fires_creek, containedby, nantahala_national_forest)
What is an active ingredient in childrens earache relief ?	(childrens_earache_relief, active_ingredients, capsicum)
What does Jimmy Neutron do?	(jimmy_neutron, fictional_character_occupation, inventor)
What dietary restriction is incompatible with kimchi?	(kimchi, incompatible_with_dietary_restrictions, veganism)

Figure: Examples of simple QA extracted from the dataset SimpleQuestions. Actual answers are underlined.

- SimpleQuestions (100k questions) [15]: contains more than 100k questions written by human annotators and associated to Freebase facts,
- WebQuestions (6k questions) is created automatically using the Google suggest API.

Open-domain vs Closed-domain QA

- The main difference between a closed-domain and an open-domain QAs is the dataset on which it was trained.
- If you train your model on the TweetQA dataset, then eventually you will get a closed-domain QA, which will be able to answer typical tweeter questions.
- Conversely, if you train a model on SQuAD, you will get a model that can answer almost any question.
- Closed-domain dataset implies that it consists of texts on a particular topic. However, a dataset with all questions in Twitter can be a closed domain, too, because such texts have a particular source. Moreover, tweets are usually stylistically specific.

Closed-domain datasets

- FICTION:
  - **CBT** is an old cloze dataset based on fiction stories for kids
  - **BookTest** encompasses all Gutenberg corpus stories
  - **FairyTaleQA** is multi-choice dataset, based on school tests
- QUIZ:
  - **TriviaQA** is based on human knowledge competitions that overlap with an encyclopedia in subject matter.
  - **Jeopardy, QuizBowl**
- REVIEWS: **AmazonQA** is based on QA and reviews on products on Amazon.com
- PROFESSIONAL: **TechQA** is a dataset of naturally occurring questions on tech expert forums
- SOCIAL NETWORK: **TweetQA** is a dataset of QAs occurring on Twitter.
- NEWS:
  - **NewsQA** is based on CNN data. Given the increasing problem of online misinformation, it is a highly important area of research, but it is hampered by the lack of public-domain data.
  - **Daily Mail Cloze** dataset is more RC-oriented

MRC vs QA

- **Machine reading comprehension** (or reading comprehension, RC) makes artificial intelligence understand a text. This "understanding" is measured by AI answering a cloze test in a dataset.
- When the machine comprehension dialog involves multiple co-referenced questions, such as if a latter question may be a logical successor of the former, the challenge is called **Conversational machine comprehension** (CMC).
- The RC task is very similar to QA, many datasets intersect (RACE, SQUAD, etc.)

Evidence	Format	Question	Answer	Example datasets
Einstein was born in 1879.	Questions	When was Einstein born?	1879	SQuAD [235], RACE [156]
	Queries	Which year Einstein born	1879	generated queries in BEIR [282]
	Cloze	Einstein was born in ___.	1979	CNN/Daily Mail [125], CBT [127]
	Completion	Einstein was born ...	in 1879	SWAG [319], RocStories [204]

Logical reasoning QA

- Logical reasoning is of vital importance to natural language understanding.
- Logical reasoning QA requires a machine to understand the logic behind the text, for example, identifying the logical components, logical relations, or fallacies.
- Some researchers propose **discourse-aware graph networks** (DAGN) to build logical graphs and learn logic representations accordingly.
- Such models analyze a context text, underlines facts, premises, and a conclusion, and then try to answer an input question.
- Let's see an example of multi-choice logical reasoning QA and the logical structure-based solution. The logical units are sentences or clauses and perform multi-hop reasoning processes from premises or refuting evidence to the conclusion.

Passage:
Astronomer: Mount Shalko is the perfect site for the proposed astronomical observatory. The summit would accommodate the complex as currently designed, with some room left for expansion. There are no large cities near the mountain, so neither smog nor artificial light interferes with atmospheric transparency. Critics claim that Mount Shalko is a unique ecological site, but the observatory need not be a threat to endemic life-forms. In fact, since it would preclude recreational use of the mountain, it should be their salvation. It is estimated that 20, 000 recreational users visit the mountain every year, posing a threat to the wildlife.

Question: Which one of the following, if true, most weakens the astronomer's argument?

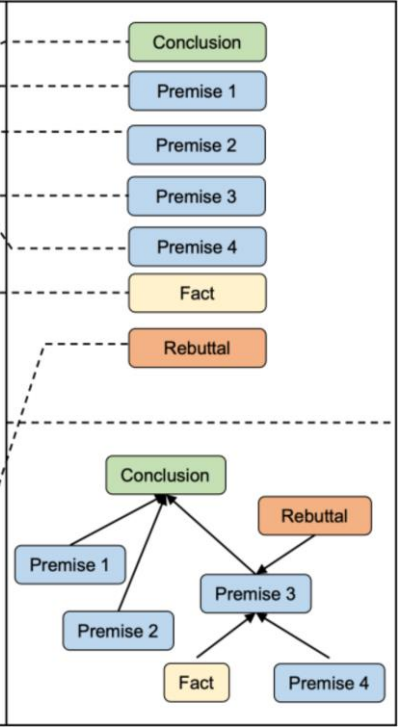
Options:

A. More than a dozen insect and plant species endemic to Mount Shalko are found nowhere else on earth.

B. The building of the observatory would not cause the small towns near Mount Shalko eventually to develop into a large city, complete with smog, bright lights, and an influx of recreation seekers.

C. A survey conducted by a team of park rangers concluded that two other mountains in the same general area have more potential for recreational use than Mount Shalko.

✓ **D.** Having a complex that covers most of the summit, as well as having the necessary security fences and access road on the mountain, could involve just as much ecological disruption as does the current level of recreational use.



Logical reasoning QA: datasets and models

- Datasets:
  - **LogiQA**, based on American Civil Servants Exams. Answers to questions from this dataset require logical reasoning. It has a text, which model should analyze, then a multi-choice question with a marked correct answer. The reasoning Type, on the far left, denotes a "degree of correctness" according to the Aristotelian logic laws.
  - **MuTual** is a retrieval-based dataset for multi-turn dialogue reasoning.
  - **ReClor**, **CLUTRR**, etc.
- Models:
  - **Rational reasoner single model** (provided by HFL & iFLYTEK). This model is based on MERIt, a MEta-path guided contrastive learning method for logical Reasoning of text.
  - **LReasoner** (based on RoBERTa), **Focal Reasoner**, and **ALBERT**.

Reasoning Type	Paragraph	Question-Answers
Categorical reasoning (30.8%)	**P1:** David knows Mr. Zhang's friend Jack, and Jack knows David's friend Ms. Lin. Everyone of them who knows Jack has a master's degree, and everyone of them who knows Ms. Lin is from Shanghai.	**Q:** Who *is from Shanghai and has a master's degree?* ✓ **A. David.** **B.** Jack. **C.** Mr. Zhang. **D.** Ms. Lin.
Sufficient conditional reasoning (27.6%)	**P2:** Jimmy asked Hank to go to the mall the next day. Hank said, "If it doesn't rain tomorrow, then I'll go climbing." The next day, there was a drizzle. Jimmy thought that Hank would not go climbing, so he went to pick up Henry to the mall. Nevertheless, Hank went climbing the mountain. When the two met again, Jimmy blamed Hank for not keeping his word.	**Q:** *Which of the following comments is appropriate?* **A.** This argument between Jimmy and Hank is meaningless. ✓ **B. Jimmy's reasoning is illogical.** **C.** The two people have different understandings of a drizzle. **D.** Hank broke his promise and caused the debate.
Necessary conditional reasoning (24.7%)	**P3:** Only if the government reinforces basic education can we improve our nation's education to a new stage. In order to stand out among other nations, we need to have a strong educational enterprise.	**Q:** *Which can be inferred from the statements above?* **A.** The whole society should be focused on education. ✓ **B. In order to stand out among nations, we should reinforce basic education.** **C.** In order to improve our education to a new stage, it is necessary to increase the salary of college teachers. **D.** In order to reinforce basic education, all primary school teachers must have a bachelor degree or above.
Disjunctive reasoning (18.5%)	**P4:** Last night, Mark either went to play in the gym or visited his teacher Tony. If Mark drove last night, he didn't go to play in the gym. Mark would go visit his teacher Tony only if he and his teacher had an appointment. In fact, Mark had no appointment with his teacher Tony in advance.	**Q:** *Which is true based on the above statements?* **A.** Mark went to the gym with his teacher Tony last night. **B.** Mark visited his teacher Tony last night. ✓ **C. Mark didn't drive last night.** **D.** Mark didn't go to the gym last night.
Conjunctive reasoning (21.3%)	**P5:**The coach of a national football team found that the best cooperative arrangement of the players U, V, W, X, Y, and Z during training are: (1) V and X can not be on the field at the same time, and neither can be off the field the same time. (2) V is not on the field only if U is not on the field. (3) If W is on the field, then X is on the field. (4) If Y and Z are on the field, then W must be on the field. This arrangement can yield the best performance.	**Q:** *If U and Z are both on the field, for best performance, which of the following arrangement is appropriate?* **A.** X is on the field and Y is not on the field. ✓ **B. V is on the field and Y is not on the field.** **C.** V and W are both on the field. **D.** V and Y are not on the field.