QA data Acquisition

Named Entity Recognition & Entity Typing

Passage Sentence

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.

Question

What causes precipitation to fall?

Answer Candidate

gravity

- Path from passage sentence words (that also occur in question) to answer.

nmod

VBZ — IN — case → NN

- Combined with path from wh-word to question word.

WP — nsubj → VBZ — advcl → VB

Text → [tokenizer → tagger → parser → ner → ...] nlp → Doc

Sentence Graph Construction & Dominating Set Derivation

x_{aug}

The **Los Angeles Lakers** are an American professional basketball team based in Los Angeles.

Question Generation

Original QA training example

Question: As of 2017, what was the estimated value of the basketball team that Luke Theodore Walton coaches?

Answer: $3.0 billion

Context: The Los Angeles Lakers are an American professional basketball team based in Los Angeles. The Lakers compete in the National Basketball Association (NBA), as a member of the league's Western Conference Pacific Division. The Lakers play their home games at Staples Center, an arena shared with the NBA's Los Angeles Clippers, the Los Angeles Sparks of the Women's National Basketball Association, and the Los Angeles Kings of the National Hockey League. The Lakers are one of the most successful teams in the history of the NBA, and have won 16 NBA championships, their last being in 2010. As of 2017, the Lakers are the second most valuable franchise in the NBA according to "Forbes", having an estimated value of $3.0 billion.
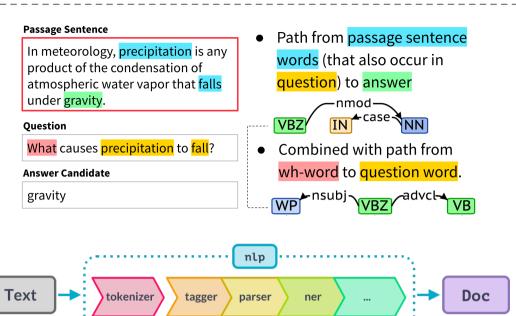
Augmented Cloze training examples

Question: What is the masked entity?

Answer: <mask>.

Context: The <mask> are an American professional basketball team based in Los Angeles. The Lakers compete in...
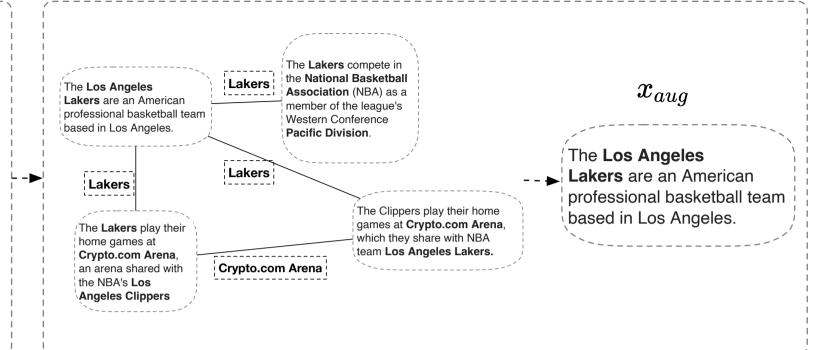
Question: What is the masked entity?

Answer: <mask>.

Context: The Los Angeles Lakers are an American professional basketball team based in <mask>. The Lakers compete in...

⋮

Prompt-style Data Augmentation

$$x^{aug} = \left[x_q^{aug} \oplus x_a \oplus x_c^{aug} \right]$$

$$y^{aug} = \left[x_q^{aug} \oplus y_a^{aug} \oplus x_c^{aug} \right]$$

(X^{aug}, Y^{aug})

$$x^{ori} = \left[x_q \oplus x_a \oplus x_c \right]$$

$$y^{ori} = \left[x_q \oplus y_a \oplus x_c \right]$$

(X^{ori}, Y^{ori})

Generative Prompt-Tuning

(X^{train}, Y^{train}) $L(\theta) = L^{ori}(\theta) + \lambda L^{aug}(\theta)$

y^{pred}

Bidirectional Encoder → Autoregressive Decoder

x^{train} $\langle s \rangle \oplus y^{pred}$