Coming Up with a Good Question Is Not Easy: A Proof

Joe Lorkowski\textsuperscript{1}, Luc Longpré\textsuperscript{1}, Olga Kosheleva\textsuperscript{2}, and Salem Benferhat\textsuperscript{3}

Departments of \textsuperscript{1}Computer Science and \textsuperscript{2}Teacher Education
University of Texas at El Paso, 500 W. University
El Paso, Texas 79968, USA
lorkowski@computer.org, longpre@utep.edu, olgak@utep.edu

\textsuperscript{3}Centre de Recherche en Informatique de Lens CRIL
Université d’Artois, F62307 Lens Cedex, France
benferhat@cril.univ-artois.fr
1. Formulation of the Problem

- Even after a very good lecture, some parts of the material remain not perfectly clear.

- A natural way to clarify these parts is to ask questions to the lecturer.

- Ideally, we should be able to ask a question that immediately clarifies the desired part of the material.

- Coming up with such good questions is a skill that takes a long time to master.

- Even for experienced people, it is not easy to come up with a question that maximally decreases uncertainty.

- In this talk, we prove that the problem of designing a good question is computationally difficult (NP-hard).
2. Towards Describing the Problem in Precise Terms: General Case

- A complete knowledge about an area means that we have the full description.
- Uncertainty means that several different variants $v_1, v_2, \ldots, v_n$ are consistent with our knowledge.
- A “yes”-‘’no” question is a question an answer to which eliminates some possible variants:
  - if the answer is “yes”, then we are limited to variants $v \in Y \subset \{v_1, \ldots, v_n\}$ consistent with “yes”;
  - if the answer is “no”, then we are limited to variants $v \in N \subset \{v_1, \ldots, v_n\}$ consistent with “no”.
- These two sets are complements to each other.
- For the question “is $v = v_1$?”, $Y = \{v_1\}$ and $N = \{v_2, \ldots, v_n\}$. 
3. Case of Probabilistic Uncertainty

- In the probabilistic approach, we assign a probability $p_i \geq 0$ to each of the possible variants: $\sum_{i=1}^{n} p_i = 1$.

- The probability $p_i$ is the frequency with which the $i$-th variant was true in similar previous situations.

- In the probabilistic case, Shannon’s entropy $S$ describes the amount of uncertainty:

$$S = - \sum_{i=1}^{n} p_i \cdot \ln(p_i).$$

- We want to select a question that maximizes the expected decrease in uncertainty.
4. How the Answer Changes the Entropy

- If the answer is “yes”, then for \( i \in N \), we get \( p'_i = 0 \), and for \( i \in Y \), we get
\[
p'_i = p(i \mid Y) = \frac{p_i}{p(Y)}, \quad \text{where } p(Y) = \sum_{i \in Y} p_i.
\]
- So, entropy changes to \( S' = -\sum_{i \in Y} p'_i \cdot \ln(p'_i) \).

- If the answer is “no”, then for \( i \in Y \), we get \( p''_i = 0 \), and for \( i \in N \), we get
\[
p''_i = p(i \mid N) = \frac{p_i}{p(N)}, \quad \text{where } p(N) = \sum_{i \in N} p_i.
\]
- So, entropy changes to \( S'' = -\sum_{i \in N} p''_i \cdot \ln(p''_i) \).

- We want to maximize the expected decrease in entropy:
\[
p(Y) \cdot (S - S') + p(N) \cdot (S - S'').
\]
5. Main Result: Probabilistic Case

- Our main result is that the problem of coming up with the best possible question is NP-hard.
- What is NP-hard: a brief reminder.
- In many real-life problems, we are looking for a string that satisfies a certain property.
- For example, in the subset sum problem:
  - we are given positive integers \( s_1, \ldots, s_n \) representing the weights, and
  - we need to divide these weights into two groups with exactly the same weight.
- So, we need to find a set \( I \subseteq \{1, \ldots, n\} \) s.t.

\[
\sum_{i \in I} s_i = \frac{1}{2} \cdot \left( \sum_{i=1}^{n} s_i \right).
\]

- The desired set $I$ can be described as a sequence of $n$ 0s and 1s: the $i$-th term is 1 if $i \in I$ and 0 if $i \notin I$.

- In principle, we can solve each such problem by simply enumerating all possible strings.

- For example, in the above case, we can try all $2^n$ possible subsets of the set \( \{1, \ldots, n\} \).

- This way, if there is a set $I$ with the desired property, we will find it.

- The problem is that for large $n$, the number $2^n$ of computational steps becomes unreasonably large.

- For example, for $n = 300$, the resulting computation time exceeds lifetime of the Universe.

- Can we solve such problems in feasible time, i.e., in time $\leq$ a polynomial of the size of the input?

- It is not known whether all exhaustive-search problems can be thus solved – this is the famous $P=\text{NP}$ problem.
- Most computer science researchers believe that some exhaustive-search problems cannot be feasibly solved.
- What is known is that some problems are the hardest (NP-hard) in the sense that
  - any exhaustive-search problem
  - can be feasibly reduced to this problem.
- This means that, unless $P=\text{NP}$, this particular problem cannot be feasibly solved.
- The above subset sum problem has been proven to be NP-hard, as well as many other similar problems.
8. How Can We Prove NP-Hardness

- A problem is NP-hard if every other exhaustive-search problem \( Q \) can be reduced to it.
- So, if we know that a problem \( P_0 \) is NP-hard, then every problem \( Q \) can be reduced to it; thus,
  - if \( P_0 \) can be reduced to our problem \( P \),
  - then, by transitivity, any problem \( Q \) can be reduced to \( P \),
  - i.e., \( P \) is indeed NP-hard.
- Thus, to prove that \( P \) is NP-hard, it is sufficient to reduce a known NP-hard problem \( P_0 \) to \( P \).
- We will prove that the subset sum problem \( P_0 \) (which is known to be NP-hard) can be reduced to \( P \).
9. Simplifying the Expression for Entropy Decrease

- For “yes”-answer, \( S' = -\sum_{i \in Y} \frac{p_i}{p(Y)} \cdot \ln \left( \frac{p_i}{p(Y)} \right) \).

- Thus, \( S' = -\frac{1}{p(Y)} \cdot \left( \sum_{i \in Y} p_i \cdot (\ln(p_i) - \ln(p(Y))) \right) \).

- So, \( S' = -\frac{1}{p(Y)} \cdot \left( \sum_{i \in Y} p_i \cdot \ln(p_i) \right) + \ln(p(Y)) \).

- Similarly, \( S'' = -\frac{1}{p(N)} \cdot \left( \sum_{i \in N} p_i \cdot \ln(p_i) \right) + \ln(p(N)) \).

- So, \( \bar{S}(Y) = p(Y) \cdot (S - S') + p(N) \cdot (S - S'') = p(Y) \cdot \ln(p(Y)) + p(N) \cdot \ln(p(N)) \).

- This expression is known to be the largest when \( p(Y) = p(N) = 0.5 \).
10. Reduction of Subset Sum to Our Problem

• Let us assume that we are given \( n \) positive integers \( s_1, \ldots, s_n \).

• Then, we can form \( n \) probabilities \( p_i \) defined as \( p_i = \frac{s_i}{\sum_{j=1}^{n} s_j} \).

• If we can find a set \( Y \) for which \( p(Y) = \sum_{i \in Y} p_i = 0.5 \), then \( \sum_{i \in Y} s_i = 0.5 \cdot \sum_{j=1}^{n} s_j \).

• This is exactly the solution to the subset sum problem.

• Vice versa, if we have a set \( Y \) for which the above equality is satisfied, then for \( p_i \) we get \( p(Y) = 0.5 \).

• The reduction shows that the problem of coming up with a good question is indeed NP-hard.
11. Case of Fuzzy Uncertainty

- In the fuzzy approach, we assign, to each variant $i$, its degree of possibility.
- The resulting fuzzy values are usually normalized, so that $\max_i \mu_i = 1$.
- One of the most widely used ways to gauge uncertainty is to use an expression $S = \sum_{i=1}^{n} f(\mu_i)$.
- Here, $f(z)$ is a strictly increasing continuous function for which $f(0) = 0$.
- This is the amount that we want to decrease by asking an appropriate question.
12. How Degrees Change After a “Yes” or “No” Answer?

- In the *numerical approach*, we normalize the remaining degree so that \( \max = 1 \), i.e., take \( \mu'_i = \frac{\mu_i}{\max_j \mu_j} \).

- In the *ordinal approach*, we raise the largest values to 1, while keeping the other values unchanged:
  \[
  \mu'_i = 1 \text{ if } \mu_i = \max_{j \in Y} \mu_j; \quad \mu'_i = \mu_i \text{ if } \mu_i < \max_{j \in Y} \mu_j.
  \]

- Based on the new values \( \mu'_i \), we compute the new complexity value \( S' = \sum_{i \in Y} f(\mu'_i) \).

- Similarly, after the “no” answer, we get \( S'' = \sum_{i \in Y} f(\mu''_i) \).

- We want to maximize the guaranteed decrease of uncertainty
  \[
  \overline{S} = \min(S - S', S - S'').
  \]
13. Main Result: Fuzzy Case

• Our main result is that the problem of coming up with the best possible question is NP-hard.

• This is true for both approaches: numerical and ordinal.

• Similarly to the probabilistic case, we prove this result by reducing the subset sum problem to this problem.

• Let $s_1, \ldots, s_m$ be positive integers.

• To solve the corresponding subset sum problem, let us:
  – select a small number $\varepsilon > 0$ and
  – consider the following $n = m + 2$ degrees:
    \[ \mu_i = f^{-1}(\varepsilon \cdot s_i) \text{ for } i \leq m \text{ and } \mu_{m+1} = \mu_{m+2} = 1. \]

- For these values $\mu_i$, we have three possible relations between the set $Y$ and the variants $m + 1$ and $m + 2$:
  1. $Y$ contains both these variants;
  2. $Y$ contains none of these two variants, and
  3. $Y$ contains exactly one of these two variants.

- Here, $f(S) = 2f(1) + O(\varepsilon)$.
  1. $f(S') = 2f(1) + O(\varepsilon)$, so $\overline{S} \leq S - S' = O(\varepsilon)$.
  2. $f(S'') = 2f(1) + O(\varepsilon)$, so $\overline{S} \leq S - S'' = O(\varepsilon)$.
  3. $f(S') = f(1) + O(\varepsilon)$ and $f(S'') = f(1) + O(\varepsilon)$, so $\overline{S} = f(1) + O(\varepsilon) \gg O(\varepsilon)$.

- Thus, the maximum of $\overline{S}$ is attained in the third case, when $\overline{S} = f(1) - \varepsilon \cdot \max \left( \sum_{i \in Y, i \leq m} s_i, \sum_{i \in N, i \leq m} s_i \right)$. 
Proof (cont-d)

- The maximum of $\overline{S}$ is attained in the third case, when
  \[
  \overline{S} = f(1) - \varepsilon \cdot \max \left( \sum_{i \in Y, i \leq m} s_i, \sum_{i \in N, i \leq m} s_i \right).
  \]
- The largest value is attained when
  \[
  \sum_{i \in Y, i \leq m} s_i = \sum_{i \in N, i \leq m} s_i = \frac{1}{2} \cdot \left( \sum_{i=1}^{m} s_i \right).
  \]
- This is exactly the solution to the subset problem.
- So, in both fuzzy approaches, the problem of coming up with a good question is indeed NP-hard.
16. What Happens in the Interval-Valued Fuzzy Case

- In many practical situations, an expert is uncertain about his/her degree of uncertainty.

- It is thus reasonable to describe the expert’s degree of certainty by a subinterval $\left[\underline{\mu}, \overline{\mu}\right] \subseteq [0, 1]$.

- Such interval-valued fuzzy techniques have indeed led to many useful applications.

- The usual fuzzy logic is a particular case of interval-valued fuzzy logic, when $\underline{\mu} = \overline{\mu}$.

- It is easy to prove that if a particular case of a problem is NP-hard, the whole problem is also NP-hard.

- Thus, the problem of selecting a good question is NP-hard for interval-valued fuzzy uncertainty as well.