Alexander Bondarenko* Yamen Ajjour Valentin Dittmar Niklas Homann Pavel Braslavski Matthias Hagen*
*first.last@informatik.uni-halle.de

Towards Understanding and Answering Comparative Questions

Motivation

- Simple comparisons: “Did Messi or Ronaldo score more goals in 2021?”
- Life-changing and highly subjective: “Is it better to move abroad or stay?”
- For big decisions, 80% of Americans rely on online research [Turner & Rainie; 2020].
- 3% of search engine’s questions are comparative [Bondarenko et al.; WSDM’20].
- 50% of these comparative questions are non-factual [Bondarenko et al.; WSDM’20].

Contributions

- Dataset: comparative questions w/ objects, aspects, answers’ stances.
- Classifiers for comparative and subjective comparative questions.
- Classifiers for direct and indirect comparative questions.
- Identifying objects, aspects, and predicates.
- Stance detector for answers.

Source: https://pixabay.com/images/id-2606759/

Comparative Questions and Answers

- 31,000 questions, 3,500 comparative, 1,690 subjective.
- 950 answers (text passages) with 4 stance labels from Stack Exchange.

Direct: `Is a cat (object) or a dog (object) a better (predicate) friend (aspect)?`

Indirect: `What pet (object) is the best (predicate) friend (aspect)?`

Without aspect: `Who is better (predicate), a cat (object) or a dog (object)?`

‘Pro cat’ answer: *Cats can be quite affectionate and attentive, and thus are good friends.*

Comparative Question Classification

- Cascading ensemble recalls 71% of comparative questions at prec. of 1.0.
  1. 10 rules: e.g., “Is a cat _or_ a dog a better_JJR friend?” Recall 54%.
  2. Feature-based: Logistic regression with word 4-grams Recall 62%.
  3. Neural: RoBERTa, BART, SBERT for representations + DNN Recall 69%.
  4. Averaging the classifiers’ decision probabilities Recall 71%.
- Operating points (probability thresholds) chosen for precision of 1.0
- Remove comparative questions after each classifiers’ group: more sophisticated classifiers for more difficult cases.
- 10-fold cross-validation.

Parsing Comparative Questions

- 10-fold cross-validation.
- Baseline: BiLSTM, 300-dimensional GloVe embeddings [Arora et al.; CIKM’17].

	F1 scores			
Classifier	**Object**	**Aspect**	**Predicate**	**None**
BiLSTM	0.80	0.52	0.85	**0.98**
RoBERTa	**0.93**	**0.80**	**0.98**	0.94

- More approaches for improving the parsing effectiveness in the paper.

Answer Stance Detection

`Is a cat (object 1) or a dog (object 2) a better (predicate) friend (aspect)?`

Pro obj. 1: *Cats can be quite affectionate and attentive, and thus are good friends.*

Pro obj. 2: *Cats are less faithful than dogs.*

- 4 labels: pro object 1, pro object 2, neutral, no stance.
- RoBERTa and Longformer for representations + DNN and logistic regression.
- RoBERTa and Longformer with sentiment prompts.
- Masking comparison objects.

`Is OBJECT 1 or OBJECT 2 a better friend?`

Pro obj. 1: *OBJECT 1 can be quite affectionate and attentive, and thus are good friends.*

- Most effective classifier RoBERTa.
- Identifying subjective questions (asking for opinions): F1 0.95.
- Comparison objects are masked in questions and answers.
- Add a sentiment prompt: OBJECT 1 *is better*.
- Input: OBJECT 1 *is better* [SEP] ANSWER.
- Highest accuracy on 4 labels (pro object 1 / 2, neutral, no stance) 0.63.

Conclusions

- Dataset: comparative questions with objects, aspects, and answers’ stances.
- Classifiers for comparative questions, objects, aspects, and predicates.
- Stance detector for potential answers.

Future Work:

- Matching comparison objects in questions and answers.
- Improving the stance detection of comparative answers.

Resources

`https://github.com/webis-de/WSDM-22`

Data: `https://webis.de/data#webis-compquestions-22`

This work has been partially supported by the Deutsche Forschungsgemeinschaft (DFG) within the project “ACQuA: Answering Comparative Questions with Arguments” that is part of the Priority Program “Robust Argumentation Machines (RATIO)”.