Contribution ID: **11** Type: **not specified**

ACQuA: Answering Comparative Questions with Arguments

In the ACQuA project, we develop algorithms to understand and answer comparative information needs like "Is a cat or a dog a better friend?" by retrieving and combining facts, opinions, and arguments from web-scale resources. Ideally, an answer explains why and under what circumstances which comparison alternative should be chosen.

Retrieval-based comparative question answering starts with identifying the important constituents: (1) the *objects* that should be compared ('cat' and 'dog' in the above example), (2) the *aspects* that indicate which properties should be emphasized in a comparative answer ('friend'), and (3) *predicates* that guide the direction of the comparison ('better'). When deriving a comparative answer by combining different sources (e.g., different web pages), the following steps can be important: (1) relevance assessment of the individual sources (e.g., a web forum on pets might be more relevant than a page on cat or dog movies), (2) quality assessment and stance detection (e.g., pro 'cat' or pro 'dog') of the retrieved arguments, (3) argument clustering based on the semantic similarity, stance, and quality, (4) re-ranking based on the predicted stance and quality, and (5) answer generation from the final ranking.

So far, our fine-tuned RoBERTa-based token classifier (trained and evaluated on 3,500 manually labeled comparative questions) can very reliably identify comparison predicates (almost perfect F1 of 0.98) and objects (F1 of 0.93), while aspect identification falls a bit behind (F1 of 0.80). Our sentiment-prompted RoBERTa-based stance detector (trained and evaluated on 950 manually labeled answers) still leaves quite some room for improvement (accuracy of 0.63). For questions that do not contain explicit objects or aspects (e.g., "What pet is best?"), we currently develop approaches that generate clarifying questions and refine the search results based on the feedback (our user study has shown that clarifying comparisons helps).

We have also developed "argumentativeness" axioms that help to re-rank documents based on (1) the number of argument units (premises and claims identified with our argument mining tool TARGER, (2) the position of query terms in the argument units, (3) (comparative) argument stance, and (4) rhetorical argument quality. Our first findings from participating in several TREC shared tasks and organizing the Touché argument retrieval shared tasks indicate that such argumentativeness facets are promising to improve rankings for argumentative information needs. However, our first results still leave room for further improvements. For instance, formulating new axioms that consider other argumentativeness facets or argument quality dimensions.

Finally, based on the aforementioned components (e.g., semantic argument similarity (argument clusters), stance, and quality), we will work on a concise abstractive answer generation / summarization from the "most relevant" arguments in the retrieved web pages. We will adapt the BiLSTM-based abstractive snippet generation framework to combine different relevant arguments into one concise answer snippet.

Primary authors: BONDARENKO, Alexander (Friedrich-Schiller-Universität Jena); HAGEN, Matthias (Friedrich-Schiller-Universität Jena)

Session Classification: Poster session