CASQAD – A New Dataset For Context-aware Spatial Question Answering

Jewgeni Rose[1][0000−0002−3053−3962] and Jens Lehmann[2,3][0000−0001−9108−4278]

[1] Volkswagen Innovation Center Europe, Wolfsburg, Germany
jewgeni.rose@volkswagen.de
[2] Computer Science III, University of Bonn, Germany
jens.lehmann@cs.uni-bonn.de
[3] Fraunhofer IAIS, Dresden, Germany
jens.lehmann@iais.fraunhofer.de
Research group: http://sda.tech

Abstract. The task of factoid question answering (QA) faces new challenges when applied in scenarios with rapidly changing context information, for example on smartphones. Instead of asking who the architect of the "Holocaust Memorial" in Berlin was, the same question could be phrased as "Who was the architect of the many stelae in front of me?" presuming the user is standing in front of it. While traditional QA systems rely on static information from knowledge bases and the analysis of named entities and predicates in the input, question answering for temporal and spatial questions imposes new challenges to the underlying methods. To tackle these challenges, we present the **C**ontext-**a**ware **S**patial **QA** **D**ataset (*CASQAD*) with over 5,000 annotated questions containing visual and spatial references that require information about the user's location and moving direction to compose a suitable query. These questions were collected in a large scale user study and annotated semi-automatically, with appropriate measures to ensure the quality.

Keywords: Datasets · Benchmark · Question Answering · Knowledge Graphs

Resource Type: Dataset
Website and documentation: http://casqad.sda.tech
Permanent URL: https://figshare.com/projects/CASQAD/80897

1 Introduction

Factoid question answering over static and massive scale knowledge bases (*KBQA*) such as DBpedia [1], Freebase [4] or YAGO [27] are well researched and recent approaches show promising performance [37]. State-of-the-art systems (e.g. [5, 11, 33, 34, 37]) perform well for simple factoid questions around a target named entity and revolving predicates. A question like *"Who was the architect of the Holocaust Memorial in Berlin, Germany?"* can be translated into a SPARQL expression to query a KB with the result *Peter Eisenman*[4]. However, in practice question answering is mostly applied in virtual

[4] He designed the Memorial to the Murdered Jews of Europe https://www.visitberlin.de/en/memorial-murdered-jews-europe

digital assistants on mobile devices, such as Siri, Alexa or Google Assistant. Users address these systems as if they are (physically) present in the situation and their communication changes compared to traditional QA scenarios. Questions contain deictic references such as “there” or “here” that need additional context information (e.g. time and geographic location [17]) to be fully understood. For example, instead of asking who the architect of the “Holocaust Memorial” in Berlin was, the same question could be phrased as *“Who was the architect of the many stelae in front of me?”* presuming the virtual assistant has knowledge about user position and viewing direction. These types of questions require the QA systems to use volatile information sets to generate the answer. Information like location or time change frequently with very different update rates. Instead of using fixed knowledge bases Context-aware QA (*CQA*) systems have to adapt to continuous information flows. That changes not only the structure of the knowledge base itself but also impacts the methodology of how to resolve the correct answer. To answer the aforementioned example question, a QA system would need an additional processing unit that provides external context with location information and matches this information set with spatial and visual signals in the input question, such as *tall* and *in front of me*.

Related works in the field of Spatial Question Answering combined geographic information system modules[5] with a semantic parsing based QA system [18,19]; proposed a system that facilitates crowdsourcing to find users that are likely nearby the according point of interest to answer temporal and location-based questions [22]; or utilizing a QA component to conduct user-friendly spatio-temporal analysis [38]. Latter is achieved by searching the input for temporal or spatial key words, which are mapped to a predefined dictionary. Despite a certain success, a commonality is that no attempt has been made to formalize and systematically combine question answering with external context, e.g. the GPS position where the question was asked. We believe, our dataset will help to close this gap, and tackling one of the main challenges in Question Answering [17].

Contribution To help bridging the gap between traditional QA systems and Context-aware QA, we offer a new and to the best of our knowledge first Context-aware Spatial QA Dataset (called *CASQAD*) focusing on questions that take spatial context information into account, i.e. visual features, user’s location and moving direction. Context has a variety of different meanings and scales, depending on application and research field. We therefore take a look at the concept of context in linguistics first and provide a crisp definition that will be used to annotate the questions. For the task of question collection, we define a case study and carry out a user study on Amazon’s MTurk crowdsourcing platform. For reproducibility, we provide the source code for the data collection and the resulting dataset[6]. In brief, our question collection and annotation process is as follows.

Raw data collection A crowdworker is presented a Human Intelligence Task (*HIT*[7]) on MTurk, containing the instructions, the project or scenario description and a StreetView

[5] A visibility engine computes, which objects are visible from the user’s point of view

[6] `https://casqad.sda.tech/`

[7] A HIT describes the micro tasks a requester posts to the workers on Amazon’s platform, also known as a “project”.

panorama embedding. The instructions cover how to control the panorama and how to pose a question with respect to our definition and goals. The scenario describes the purpose and background for the case study, which is as follows: "Imagine driving through a foreign city and ask questions about the surrounding you would usually ask a local guide". The panorama is a StreetView HTML embedding that the user can rotate and zoom in or out but not move freely around the streets, which forces the focus on the presented points of interest. To ensure the quality of the collected questions, we developed comprehensive guidelines, including splitting the batches and monitoring the collection process. We collected over 5,000 questions in sum for 25 panoramas in the German city of Hanover from over 400 different workers.

Question annotation For the annotation process we follow a two-step approach. First, we pre-process the raw input automatically to detect named entities, spatial and visual signals[8] and annotate the questions. Second, three human operators evaluate these question-annotation pairs and either approve or correct them.

2 Related Work

In recent years multiple new datasets have been published for the task of QA [2, 3, 6, 9, 13–15,20,21,24,25,28,35], including benchmarks provided by the Question Answering over Linked Data (QALD) challenge [29–31]. The datasets and benchmarks differ particularly in size (a few hundreds to hundreds of thousands), complexity (simple facts vs. compositional questions), naturalness (artificially generated from KB triples vs. manually created by human experts), language (mono- vs. multilingual) and the underlying knowledge base (DBpedia, YAGO2, Freebase or Wikidata), in case SPARQL queries are provided.

SimpleQuestions [6] and WebQuestions [3] are the most popular datasets for the evaluation of simple factoid question answering systems, despite the fact that most of the questions can already be answered by standard methods [12, 23]. The recent QALD benchmarks contain more complex questions of higher quality with aggregations and additional filter conditions, such as "Name all buildings in London which are higher than 100m" [31]. These questions are hand-written by the organizers of the challenge and are small in number (up to a few hundred questions).

The SQuAD [25] dataset introduces 100,000 crowdsourced questions for the reading comprehension task. The crowdworkers formulate a question after reading a short text snippet from Wikipedia that contains the answer. The SQuAD 2.0 [24] dataset introduces unanswerable questions to make the systems more robust by penalizing approaches that heavily rely on type matching heuristics. NarrativeQA [21] presents questions which require deep reasoning to understand the narrative of a text rather than matching the question to a short text snippet. The recently published LC-QuAD 2.0 [14] dataset contains 30,000 questions, their paraphrases and corresponding SPARQL queries. The questions were collected by verbalizing SPARQL queries that are generated based on hand-written templates around selected entities and predicates. These verbalizations

[8] Using state-of-the-art models from `https://spacy.io/`

are then corrected and paraphrased by crowdworkers. For a more detailed description and comparison of standard and recent benchmarks for question answering, we refer to [32, 36].

The TempQuestions [20] benchmark contains 1,271 questions with a focus on the temporal dimension in question answering, such as *"Which actress starred in Besson's first science fiction and later married him?"*, which requires changes to the underlying methods regarding question decomposition and reasoning about time points and intervals [20]. The questions were selected from three publicly available benchmarks [2,3,9] by applying hand-crafted rules and patterns that fit the definition of temporal questions, and verified by human operators in the post-processing. However, the processing of questions specifically containing spatial or visual references that require additional context information to be answered was not considered so far.

3 Defining Spatial Questions

There are various different types of questions, which require additional information to be fully understood. Questions can contain a personal aspect, cultural background, or simply visual references to the surrounding location. More formally, in linguistics context is described as a frame that surrounds a (focal) event being examined and provides resources for its appropriate interpretation [8]. This concept is extended by four dimensions, namely setting, behavioral environment, language and extra-situational context. Behavioral environment and language describe how a person speaks and how she presents herself, i.e. the use of gestures, facial expressions, speech emphasis or use of specific words. For instance, this can be used to differentiate between literally or sarcastically meant phrases. The setting describes the social and spatial framework and the extra-situational dimension provides deeper background knowledge about the participants, e.g. the personal relationship and where a conversation is actually held (office vs. home). All dimensions describe important information to process a question properly. To make a first step towards Context-aware QA, in this work we focus on the setting dimension, specifically questions containing spatial and visual references, which require reasoning over multiple data sources. A spatial location is defined by its 2-dimensional geo-coordinate (latitude and longitude). However, users in a real-world scenario rather ask for information about a target object by referring or relate to visually more salient adjacent objects or describe the target visually or both. For this reason, we will define a spatial question by the visual and spatial signals contained in the phrase. The task of spatial question collection is covered in Section 4.

3.1 Spatial Signals

We refer to spatial signals as keywords or phrases that modify a question such that it requires a QA system to have additional knowledge about the spatial surrounding of the user. Table 1 shows samples of spatial signals used in the context of spatial question answering applied on mobile assistants. Deictic references are used to point to entities without knowing the name or label, such as *that* building. Positional or vicinity signals reinforce the disambiguation of nearby entities by facilitating the matching between the input question and possible surrounding entities. For example, in the question “What is the column next to the Spanish restaurant?” the signal *next to* is used to point to the column that is next to the more salient object “Spanish restaurant”.

Table 1: Spatial signals examples distributed over different categories with according text snippets. Further, all spatial signals can be combined, such as in “What is *that* building *next to* the book store?”

	Spatial Signals	**Snippet**
Deixis	That, This, There	“over there”
Position	Left, right, in front	“left to me”
Vicinity	Next to, after, at	“right next to the book store”

3.2 Visual Signals

Visual signals are keywords and phrases that specify or filter the questions target entity. Similar to the spatial signals for position and vicinity they facilitate the disambiguation of nearby entities or entities in the same direction from the user’s point of view. Visual signals are stronger in terms of filtering visible salient features and attributes, such as color, shape or unique features. Table 2 shows samples of visual signals for different categories.

Table 2: Visual signal examples for different categories with according example snippets.

	Visual Signals	**Snippet**
Color	Red, green, blue	“yellowish building”
Size	Tall, small, big, long	“tall column”
Shape	Flat, rounded, conical	“rounded corners”
Salience	Flags, brick wall, glass	“flags on the roof”

3.3 Spatial Questions

Utilizing the described concepts for spatial and visual signals from the Sections 3.1 and 3.2 we can define a spatial question as follows:

Definition 1. *A spatial question contains at least one spatial signal and requires additional context knowledge to understand the question and disambiguate the target entity. A spatial question can contain multiple visual signals.*

The results of our theoretical considerations disclose challenges to QA systems dealing with spatial questions corresponding to our definition. The QA system requires additional knowledge about user's position, moving or viewing direction and surroundings. The questions contain deictic references (*that*) and location information (*next to*), making it impossible to use traditional approaches based on named entity recognition. The exemplary question taken from our case study that will be presented in Section 4, shows the need for new methods for CQA.

Example 1. "What is the white building on the corner with the flags out front?"

Here, we have visual signals *white building* (color) and *with flags out front* (salience) which filter the possible entity candidates for the spatial signal *on the corner*. These filters are important to pinpoint the target entity with a higher probability. Even with distinctive spatial signals such as *on the corner*, we could face four different buildings to choose from – potentially even more, in case there are multiple buildings.

4 CASQAD – Context-aware Spatial QA Dataset

The main objective of our work is the introduction of a spatial questions corpus that fits the definition in Section 3.3, i.e. the questions require the QA system to combine different input sources (at least one for the question and one for the context information) to reason about the question objective and target entity, which is a big step towards Context-aware Question Answering. The most intuitive way to collect natural questions with minimized bias, is to conduct a user study. We use Amazon's MTurk crowdsourcing platform for this task, considering common best practices [10]. MTurk is an efficient option to collect data in a timely fashion [7] and is the de facto standard to collect human generated data for natural language processing. Since we are interested in spatial questions we have to design the collection task accordingly. Therefore, we focus on our motivational scenario that pictures the use of a Context-aware QA system on a mobile device.

4.1 Experimental Setup on MTurk

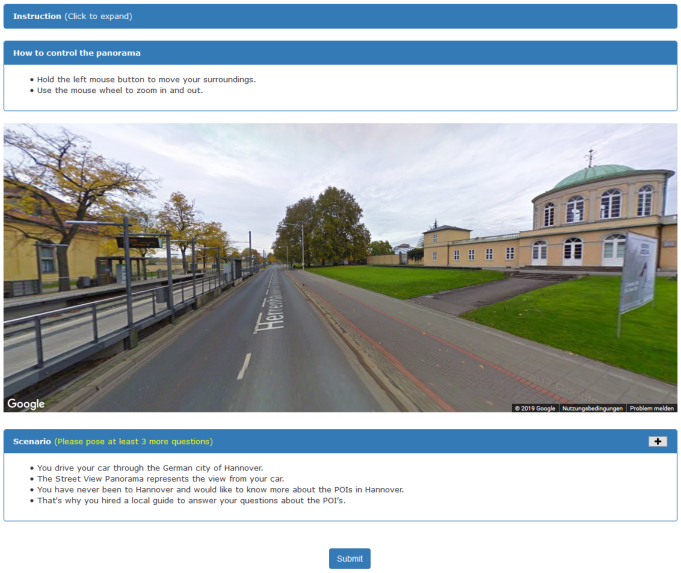
For the collection task, we first define an appropriate scenario and design to instruct the crowdworkers at MTurk (also called *turkers*). Instead of showing a textual description from Wikipedia containing the answer, we present the task in a more natural way, which also fits our scenario. We show a Google Street View[9] HTML embedding in the survey with the following instructions: "Imagine driving through the German city of Hanover, which is foreign to you. To get to know the city better, you hire a local guide who can answer your questions about surrounding points of interest (POIs). The Street View panorama represents the view from your car." Street View is used on MTurk for various image annotation tasks, for example to support the development of vision-based driver assistance systems [26]. Hara et al. [16] incorporated Street View images in a MTurk survey to identify street-level accessibility problems. In contrast to static images we embed dynamic Street View panoramas in an HTML document, which facilitates an interactive user-system interaction.

[9] `https://www.google.com/intl/en/streetview/`

Instructions:

- Please ask questions about surrounding POIs visible in the panorama.
- When posing a question, please make sure the current field of view is oriented towards the questions subject.
- General questions such as “Where am I?” will not be awarded.
- You drive through the German city of Hanover.
- The Street View Panorama represents the view from your car.
- You have never been to Hanover and would like to know more about the POIs.
- That’s why you hired a local guide to answer your questions about the POI’s.

The Route: A crucial part to ensure validity of the experiments is the choice of anchor points for the Street View panoramas. An anchor point is the initial point of view that is presented to the turker. Our goal is to collect spatial questions that people would ask about visible surroundings. For this reason, we picked panoramas containing several POIs from a typical commercial tourist city tour in Hanover[10]. The route consists of 24 different panoramas showing 521 directly visible POIs (buildings, stations, monuments, parks) that have an entry in Open Street Map[11]. Further, some of the panoramas show dynamic objects that were present at the time the pictures were taken, such as pedestrians and vehicles.



Instruction (Click to expand)

How to control the panorama

- Hold the left mouse button to move your surroundings.
- Use the mouse wheel to zoom in and out.

Scenario (Please pose at least 3 more questions)

- You drive your car through the German city of Hannover.
- The Street View Panorama represents the view from your car.
- You have never been to Hannover and would like to know more about the POIs in Hannover.
- That's why you hired a local guide to answer your questions about the POI's.

Fig. 1: An example Street View embedding showing the train stop *Hannover Herrenhäuser Gärten* to the left and the *Library Pavilion* with the *Berggarten* to the right.

[10] `https://www.visit-hannover.com/en/Sightseeing-City-Tours/Sightseeing/City-tours`

[11] `https://www.openstreetmap.org`

Table 3: The table shows all 24 anchor points from the route including the position (longitude and latitude), the heading as angle degree (0° is north, 90° is east, 180° is south, and 270° is west), and the number of visible points of interest (POIs) and buildings (e.g. office or apartment buildings).

Title	**Latitude**	**Longitude**	**Heading**	**Visible POIs**	**Visible Buildings**
Schlosshäuser im Berggarten Hannover	52.3915561	9.7001302	0	8	24
Landesmuseum Hannover und Staatskanzlei Niedersachsen	52.365528	9.7418318	228	20	24
Neues Rathaus	52.3680496	9.7366261	169	10	14
Marktkirche Hannover	52.3722614	9.7353854	188	12	27
Staatstheater Hannover	52.3737913	9.7417629	238	8	31
Landesmuseum Hannover	52.3650037	9.739624	20	23	11
Leibnitz Universität	52.3816144	9.7175591	7	19	12
Musiktheater Bahnhof Leinhausen	52.3963175	9.6770442	8	8	12
Stöckener Friedhof	52.4003496	9.6692401	11	27	28
Marktkirche Hannover	52.372351	9.7352942	194	12	27
Christuskirche (Hannover)	52.3816745	9.7262545	198	9	58
Neues Rathaus	52.3677048	9.7386612	240	15	23
Landesmuseum	52.3655038	9.7397131	93	17	6
Döhrener Turm	52.3467714	9.7605805	5	19	22
Amtsgericht	52.3773252	9.7449014	168	3	5
VW-Tower	52.3798545	9.7419755	285	15	20
Niedersächsisches Finanzministerium	52.3723053	9.7455448	134	3	35
Börse Hannover	52.3723032	9.7417374	127	11	29
Niedersächsisches Wirtschaftsministerium	52.3689436	9.7347076	29	19	16
Ruine der Aegidienkirche	52.3692021	9.738762	64	11	30
Waterloosäule	52.3663441	9.726473	78	8	19
Niedersachsenhalle	52.3770004	9.7693496	198	17	24
Landtag Niedersachsen	52.3707776	9.7336929	218	6	12
Hauptbahnhof	52.3759631	9.7401624	0	24	18

The HITs: A Human Intelligence Task (HIT) describes the task a crowdworker is supposed to solve in order to earn the reward. The requester has to provide information for the worker including a (unique) title, job description and clear reward conditions. In addition the requester has to specify qualification requirements in the MTurk form to filter desirable from undesirable crowdworkers, such as gender, age, profession, or more specific qualifications like having a driving license or visited places. Here, we specified crowdworkers to be equally distributed over the common age groups and gender. All workers are English speakers and have their residence in the United States, spread proportionally among the population of the individual states[12]. Additionally, we asked the workers if they ever visited the German city of Hanover before, to make sure the scenario of visiting a foreign city holds to minimize the bias. The task for the turkers is to pose at least three different questions to the system, which shows one of the 24 panoramas. When the user submits a question, she has to focus the view on the target object, e.g. the bridge or monument. As a result, we automatically annotate the question with potential context information, by analyzing the position, viewing direction, pitch and zoom level of the panorama[13]. To prevent empty or too short questions, we analyze the input in real time. This is achieved by hosting the web page, which is embedded into the MTurk form, on our own servers on Azure. In our experiments every turker is limited up to eight HITs, which is a good trade-off between cost efficiency and diversity of the workers.

The MTurk form: Figure 1 shows a screenshot from the document presented to the turkers. On top of the figure is the collapsible instructions box with general instructions, for example, what button to click to submit a question, and reward constraints (the turkers are not paid, if we detect spam, fraud attempts or any random input). To lessen the distraction in the view, we don't use control panels in the embedding. There is a small panel with a control description above the Street View embedding, which is the center of the form. Using the mouse, the turker has a 360 degree view and can change the pitch and zoom level. In contrast to Google's web application, the turkers cannot move freely around in the panorama, i.e. change their position. This is done to force the focus of the turker on the visible objects in the given panorama. After completing a HIT, randomly another HIT containing a different anchor point from the route is offered to the turker. A panorama is never presented twice to the same user.

4.2 Annotation Process

In sum we collected 5,232 valid[14] questions by 472 different turkers. An exploratory analysis shows, that the questions range from questions about salient buildings, like "Is this a government building?" to questions such as "Is the bus station a good spot to pick up girls?". However, most of the questions are about nice places to stay and eat or interesting looking monuments nearby – questions to be expected from a foreigner to ask an

[12] `https://en.wikipedia.org/wiki/List_of_states_and_territories_of_the_United_States_by_population`

[13] All meta information is provided by Google's Street View API `https://developers.google.com/maps/documentation/streetview/`

[14] We removed manually questions such as "Who am I?"

assistant or to look up in a city guide. More specific, the turkers asked for information about the cuisine and opening hours of nearby restaurants and theaters, or building dates and architectural styles. The required information to answer these questions is typically available in common knowledge sources such as OpenStreetMap[15], Google Places[16] or Wikidata[17]. More details about the annotation analysis will be presented in Section 4.4. We annotated the dataset in a two-step approach that will be presented below.

Automated Processing: The first processing step is normalizing the user input. Sentences containing multiple questions are separated and white space characters normalized first[18]. For example *"What is this building? When can I visit it?"* is separated into two questions. Then, every question is labeled automatically with relevant meta data from Street View, i.e. position, heading, pitch and zoom level, and the according Street View panorama direct HTTP link. Storing and sharing the images is not permitted per terms of use.

Manual Processing: In the second processing step, three local experts within our team annotate the previously processed questions. We prepared a form containing the raw input, the normalized questions, the meta information from Google Street View and the according image. The annotation task was to tag the questions with the objective of the question, such as the age of a building, mark vicinity phrases as explicit spatial references, as well as phrases containing visual signals. We differentiate between vicinity and simple deictic references to express the complexity and difficulty of these questions, such as *"What is across the street from the Borse building?"*. Finally, the annotators have to choose the questions target object, such as a POI, a nearby location (*"Is **this area** safe at night?"*) or something else (e.g. questions such as *"In what direction is the capitol?"*).

4.3 Experiments with Crowd-based Annotations

In an early experiment with a batch of 200 Hits we attempted to annotate the phrases by the crowdworkers. We created an additional input mask in the MTurk data collection questionnaire, in which the crowdworkers were supposed to annotate their questions themselves. Using the meta information provided by the Google Street View API we approximated the visible objects in a panorama, queried every available information in the aforementioned knowledge sources (Google Places, OSM and Wikidata) and offered a list of possible answers or information for all records. Then we asked the turkers to annotate the questions with the following information:

1. Choose the object of interest from the given list of objects (object displayed including name, type, and a list of all available attributes)
2. Choose the intent of your question (this is basically a record from the list of attributes, such as construction date for buildings or cuisine for restaurants)
3. If there is no appropriate entry, choose "misc" for object or intent

[15] `https://www.openstreetmap.org`

[16] `https://cloud.google.com/maps-platform/places`

[17] `https://www.wikidata.org/`

[18] Even though we instructed the turkers to phrase only one question per input frame, not all followed the instruction.

However, our evaluation revealed critical flaws in this process. The number of approximated visible objects was too high for each panorama to disambiguate these correctly, especially for non-locals. Consequently, the same applies for the choice of the right intent. In addition, the missing English terms for German local places made it difficult to understand the meaning or usage of a place or building. The error rate was over 50% (not including the cases when the crowdworkers selected "misc" as the intent or object). We decided not to use the crowdworker annotations, if every annotation had to be checked by experts again anyway.

4.4 Annotation Analysis

The questions length ranges from 3 to 31 words, whereas the average length of the words is 4.4. The average number of words per question is 6.8 and the according median is 6, which is similar to comparable datasets [20]. Table 4 shows the frequency count for the first token of the question at the left columns. The right columns show the frequency count for the question intent. Both lists cover similarly 84% of all questions. Figure 2 shows the comparison of the word distribution with related datasets.

Table 4: Top 10 list with first token and intent frequency count.

First token	**Count**	**Intent**	**Count**
What	2368	Category	1288
is	1017	Construction Date	671
how	502	Name	292
when	414	Usage	281
are	176	Opening Hours	144
do(es)	137	Significance	120
can	113	Offering	116
who	105	Accessibility	91
where	101	Architecture	81
	2952		**3084**

Spatial and Visual Signals: A detailed analysis shows, that the turkers phrase 92% of the questions using simple deictic references to refer to nearby points of interest, otherwise naming the entities (e.g. some businesses have a name on the entrance sign). Questions that contain explicit signals for vicinity or visual information are less frequent. On the other hand, these questions are more complex and challenging, for example *"What is behind the field across from the large building?"*. Table 5 shows the distribution of spatial and visual signals of the annotated questions.

5 Conclusion

We published a new dataset containing 5,232 textual questions that are spatial by nature – *CASQAD*. In addition we enriched the questions with according meta context

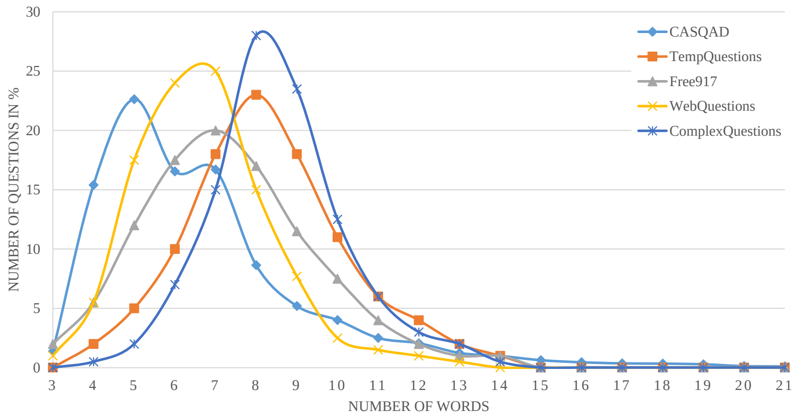


Fig. 2: Distribution of words per question in CASQAD compared to some popular datasets for Question Answering over Knowledge Bases: ComplexQuestions [2], WebQuestions [3], Free917 [9], and TempQuestions [20]

Table 5: Questions distribution by spatial signals. Questions containing named entities usually aren't spatial by our definition.

Signal Type	Example Question	Total
Visual	*What's inside the large stone building?*	490
Vicinity	*Are there any good pubs around here?*	350
Deixis	*What type of architecture is this?*	4839
Size	*How tall is this building?*	260
Color	*What is the building over there with the blue symbol?*	214
Named Entity	*What happens at the Amt-G. Hannover?*	402

information from Google Street View, such as the GPS position of the point of view, and direct links to the according images. The questions complexity ranges from rather simple questions querying one attribute of a point of interest, to questions about the social and historical background of specific symbols in the images. The versatility of this dataset facilitates the usage for KBQA as well as for text comprehension, or hybrid systems including visual QA. We hope to spur research Context-aware Question Answering systems with this dataset. *CASQAD* is currently being used in multiple internal projects in the Volkswagen Group Innovation[19], in particular in the research field of digital assistants. Future work will include a ready to use end-to-end baseline and an extensive evaluation in a real world scenario with users driving in a car to explore a foreign city, to encourage further research in this direction.

[19] `https://www.volkswagenag.com/en/group/research---innovations.html`

References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: A Nucleus for a Web of Open Data. In: The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, vol. 4825 LNCS, pp. 722–735 (2007). https://doi.org/10.1007/978-3-540-76298-0_52
2. Bao, J., Duan, N., Yan, Z., Zhou, M., Zhao, T.: Constraint-Based Question Answering with Knowledge Graph. In: Proceedings of {COLING} 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 2503–2514. The COLING 2016 Organizing Committee, Osaka, Japan (2016)
3. Berant, J., Chou, A., Frostig, R., Liang, P.: Semantic Parsing on Freebase from Question-Answer Pairs. Proceedings of EMNLP (October), 1533–1544 (2013)
4. Bollacker, K., Cook, R., Tufts, P.: Freebase: A shared database of structured general human knowledge. Proceedings of the national conference on Artificial Intelligence **22**(2), 1962 (2007)
5. Bordes, A., Chopra, S., Weston, J.: Question Answering with Subgraph Embeddings. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (October 25-29), 615–620 (2014). https://doi.org/10.3115/v1/D14-1067
6. Bordes, A., Usunier, N., Chopra, S., Weston, J.: Large-scale Simple Question Answering with Memory Networks (2015). https://doi.org/10.1016/j.geomphys.2016.04.013
7. Buhrmester, M.D., Talaifar, S., Gosling, S.D.: An Evaluation of Amazon's Mechanical Turk, Its Rapid Rise, and Its Effective Use. Perspectives on Psychological Science **13**(2), 149–154 (2018). https://doi.org/10.1177/1745691617706516
8. Bulcaen, C.: Rethinking Context: Language as an Interactive Phenomenon. Language and Literature **4**(1), 61–64 (feb 1995). https://doi.org/10.1177/096394709500400105
9. Cai, Q., Yates, A.: Large-scale Semantic Parsing via Schema Matching and Lexicon Extension. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 423–433. Association for Computational Linguistics, Sofia, Bulgaria (aug 2013)
10. Cheung, J.H., Burns, D.K., Sinclair, R.R., Sliter, M.: Amazon Mechanical Turk in Organizational Psychology: An Evaluation and Practical Recommendations. Journal of Business and Psychology **32**(4), 347–361 (2017). https://doi.org/10.1007/s10869-016-9458-5
11. Dhingra, B., Danish, D., Rajagopal, D.: Simple and Effective Semi-Supervised Question Answering. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). pp. 582–587. Association for Computational Linguistics, Stroudsburg, PA, USA (2018). https://doi.org/10.18653/v1/N18-2092
12. Diefenbach, D., Lopez, V., Singh, K., Maret, P.: Core techniques of question answering systems over knowledge bases: a survey. Knowledge and Information Systems (October), 1–41 (sep 2017). https://doi.org/10.1007/s10115-017-1100-y
13. Diefenbach, D., Tanon, T.P., Singh, K., Maret, P.: Question answering benchmarks for Wikidata. CEUR Workshop Proceedings **1963**, 3–6 (2017)
14. Dubey, M., Banerjee, D., Abdelkawi, A.: LC-QuAD 2 . 0 : A Large Dataset for Complex Question Answering over Wikidata and DBpedia
15. Dunn, M., Sagun, L., Higgins, M., Guney, V.U., Cirik, V., Cho, K.: SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine (2017)
16. Hara, K., Le, V., Froehlich, J.: Combining crowdsourcing and google street view to identify street-level accessibility problems. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13 p. 631 (2013). https://doi.org/10.1145/2470654.2470744

17. Höffner, K., Walter, S., Marx, E., Usbeck, R., Lehmann, J., Ngonga Ngomo, A.C.: Survey on challenges of Question Answering in the Semantic Web. Semantic Web **8**(6), 895–920 (2017). https://doi.org/10.3233/SW-160247
18. Janarthanam, S., Lemon, O., Bartie, P., Dalmas, T., Dickinson, A., Liu, X., Mackaness, W., Webber, B.: Evaluating a City Exploration Dialogue System Combining Question-Answering and Pedestrian Navigation. 51st Annual Meeting of the Association of Computational Linguistics (October 2015), 1660–1668 (2013). https://doi.org/10.18411/a-2017-023
19. Janarthanam, S., Lemon, O., Liu, X., Bartie, P., Mackaness, W., Dalmas, T., Goetze, J.: Integrating Location, Visibility, and Question-Answering in a Spoken Dialogue System for Pedestrian City Exploration. Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (July), 134–136 (2012)
20. Jia, Z., Abujabal, A., Roy, R.S., Strötgen, J., Weikum, G.: TempQuestions: {A} Benchmark for Temporal Question Answering. {WWW} (Companion Volume) **2**, 1057–1062 (2018)
21. Kočiský, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K.M., Melis, G., Grefenstette, E.: The NarrativeQA Reading Comprehension Challenge. Transactions of the Association for Computational Linguistics **6**, 317–328 (2018). https://doi.org/10.1162/tacl_a_00023
22. Liu, Y., Alexandrova, T., Nakajima, T.: Using stranger [sic] as sensors: temporal and geo-sensitive question answering via social media. WWW '13: Proceedings of the 22nd international conference on World Wide Web pp. 803–813 (2013). https://doi.org/10.1145/2488388.2488458
23. Petrochuk, M., Zettlemoyer, L.: SimpleQuestions Nearly Solved: A New Upperbound and Baseline Approach. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 554–558. Association for Computational Linguistics, Stroudsburg, PA, USA (2018). https://doi.org/10.18653/v1/D18-1051
24. Rajpurkar, P., Jia, R., Liang, P.: Know what you don't know: Unanswerable questions for SQuAD. ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers) **2**, 784–789 (2018)
25. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ Questions for Machine Comprehension of Text. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (ii), 2383–2392 (2016). https://doi.org/10.18653/v1/D16-1264
26. Salmen, J., Houben, S., Schlipsing, M.: Google street view images support the development of vision-based driver assistance systems. Proceedings of the IEEE Intelligent Vehicles Symposium (June 2012), 891–895 (2012). https://doi.org/10.1109/IVS.2012.6232195
27. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A Core of Semantic Knowledge. In: Proceedings of the 16th International Conference on World Wide Web (2007). https://doi.org/10.1145/1242572.1242667
28. Trivedi, P., Maheshwari, G., Dubey, M., Lehmann, J.: LC-QuAD: A corpus for complex question answering over knowledge graphs. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **10588 LNCS**, 210–218 (2017). https://doi.org/10.1007/978-3-319-68204-4_22
29. Unger, C., Forascu, C., Lopez, V., Ngomo, A.C.N., Cabrio, E., Cimiano, P., Walter, S.: Question answering over linked data (QALD-4). CEUR Workshop Proceedings **1180**, 1172–1180 (2014)
30. Unger, C., Forascu, C., Lopez, V., Ngomo, A.C.N., Cabrio, E., Cimiano, P., Walter, S.: Question answering over linked data (QALD-5). CLEF **1180**, 1172–1180 (2015)
31. Usbeck, R., Ngomo, A.C.N., Haarmann, B., Krithara, A., Röder, M., Napolitano, G.: 7th open challenge on question answering over linked data (QALD-7). In: Communications in Computer and Information Science, vol. 769, pp. 59–69 (2017). https://doi.org/10.1007/978-3-319-69146-6_6

32. Usbeck, R., Röder, M., Hoffmann, M., Conrads, F., Huthmann, J., Ngonga-Ngomo, A.C., Demmler, C., Unger, C.: Benchmarking Question Answering Systems. Semantic Web **1**, 1–5 (2016)
33. Yang, M.C., Duan, N., Zhou, M., Rim, H.C.: Joint Relational Embeddings for Knowledge-based Question Answering. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 645–650 (2014). https://doi.org/10.3115/v1/D14-1071
34. Yang, Z., Hu, J., Salakhutdinov, R., Cohen, W.: Semi-Supervised QA with Generative Domain-Adaptive Nets. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1040–1050. Association for Computational Linguistics, Stroudsburg, PA, USA (2017). https://doi.org/10.18653/v1/P17-1096
35. Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., Manning, C.D.: HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2369–2380. Association for Computational Linguistics, Stroudsburg, PA, USA (2019). https://doi.org/10.18653/v1/D18-1259
36. Yatskar, M.: A Qualitative Comparison of CoQA, SQuAD 2.0 and QuAC. NAACL-HLT (sep 2018)
37. Yih, W.T., Chang, M.W., He, X., Gao, J.: Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base. Acl pp. 1321–1331 (2015). https://doi.org/10.3115/v1/P15-1128
38. Yin, Z., Goldberg, D.W., Zhang, C., Prasad, S.: An NLP-based question answering framework for spatio-temporal analysis and visualization. ACM International Conference Proceeding Series **Part F1482**, 61–65 (2019). https://doi.org/10.1145/3318236.3318240