Joint Spatio-Textual Reasoning for Answering Tourism Questions

Danish Contractor*
IBM Research AI & Indian Institute of Technology – Delhi
New Delhi, India
dcontrac@in.ibm.com

Shashank Goel
Indian Institute of Technology – Delhi
New Delhi, India
shashankgoel.iitd@gmail.com

Mausam
Indian Institute of Technology – Delhi
New Delhi, India
mausam@cse.iitd.ac.in

Parag Singla
Indian Institute of Technology – Delhi
New Delhi, India
parags@cse.iitd.ac.in

ABSTRACT

Our goal is to answer real-world tourism questions that seek Points-of-Interest (POI) recommendations. Such questions express various kinds of spatial and non-spatial constraints, necessitating a combination of textual and spatial reasoning. In response, we develop the first joint spatio-textual reasoning model, which combines geo-spatial knowledge with information in textual corpora to answer questions. We first develop a modular spatial-reasoning network that uses geo-coordinates of location names mentioned in a question, and of candidate answer POIs, to reason over only spatial constraints. We then combine our spatial-reasoner with a textual reasoner in a joint model and present experiments on a real world POI recommendation task. We report substantial improvements over existing models without joint spatio-textual reasoning. To the best of our knowledge, we are the first to develop a joint QA model that combines reasoning over external geo-spatial knowledge along with textual reasoning.

CCS CONCEPTS

• **Information systems** → **Question answering**.

KEYWORDS

spatio-textual reasoning, joint model, POI, question answering, web data, entity reviews

ACM Reference Format:

Danish Contractor, Shashank Goel, Mausam, and Parag Singla. 2021. Joint Spatio-Textual Reasoning for Answering Tourism Questions. In *Proceedings of the Web Conference 2021 (WWW '21), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3442381.3449857

1 INTRODUCTION

Users of travel forums often post questions seeking personalized recommendations for their travel needs. Consider the example in Figure 1, which shows a real-world[1] Points-of-Interest (POI) seeking question. Answering such a recommendation question is a challenging problem as, it not only requires reasoning over a text corpus describing potential restaurants (eg. reviews), but it also requires resolving spatial constraints ("near Hotel Florida") over the physical location of a restaurant. In addition, the question is also under-specified and ambiguous (eg, "dont have to venture too far") making the spatial-inference task harder.

Question: *"We're arriving into Havana from the UK in the late afternoon and are staying at the Hotel Florida. By the time we get there and check in it will be early evening. Can anyone recommend a good restaurant nearby so that we dot have to venture too far on our first night after a long journey? The Hotel is at Calle Obispoesq. a Cuba. Ciudad de La Habana. Any suggestions appreciated."*
Answers:
1. `Los Nardos' (39_R_2966)
2. 'La Mina' (39_R_2303)

Figure 1: A sample POI recommendation question. The answers correspond to POI IDs of the form <city_id >_<POI type>_<number>. The Tourism QA dataset [9] has three classes of POIs - restaurants (R), attractions (A) and hotels (H).

Recently, there has been work on QA models that fuse knowledge from multiple sources; for example, by combining data from knowledge bases with textual passages [2, 37], or incorporating multi-modal data sources [16, 35]. But, we do not know of systems that fuse geo-spatial knowledge with text. In addition, there exist several geo-spatial IR systems (eg, [29, 30]), however, to the best of our knowledge, none of them perform joint-reasoning over geo-spatial and textual knowledge sources.

In response, we present our joint spatio-textual QA model for returning answers to questions that require textual as well as spatial reasoning. We first develop a modular spatial-reasoning network that uses geo-coordinates of location names mentioned in a question, and, of candidate answer entities, to reason over only spatial constraints. It learns to associate contextual *distance-weights* with each location-mention in the question – these weights are combined

*The author is a PhD student at IIT Delhi and an employee of IBM Research.

[1] https://bit.ly/2zIxQpj

WWW '21, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8312-7/21/04.
https://doi.org/10.1145/3442381.3449857

with their respective spatial-distances from a candidate answer, to generate a 'spatial relevance' score for that answer.

We then combine the spatial-reasoner with a textual QA system to develop a joint spatio-textual[2] QA model. We demonstrate the model using a recently introduced QA task, which contains tourism questions seeking POI (entity) answers [9]. It also contains a collection of entity reviews as knowledge source for answering these questions. We provide the geo-spatial knowledge for the task by mapping location-mentions in questions to their geographical coordinates using publicly available APIs. Similarly, candidate answer POIs are also mapped to their geographical coordinates, included as part of the dataset [9]. To the best of our knowledge, we are the first to develop a joint QA model that combines reasoning over external geo-spatial knowledge along with textual reasoning.

Contributions: Our paper makes the following contributions:

1. We develop a spatial-reasoner that uses geo-coordinates of locations and POIs to reason over spatial constraints specified in a question.
2. We demonstrate, using a simple synthetic dataset, that our spatial-reasoner is not only able to reason over "near", "far" constraints but is also able to determine location references that are not useful for reasoning (Eg: a location reference mentioning where a user last went on vacation).
3. We develop a spatio-textual QA model, which fuses spatial knowledge (geo-coordinates) with textual knowledge (POI reviews) using sub-networks designed for spatial and textual reasoning.
4. We demonstrate that our joint spatio-textual model performs significantly better than models employing only spatial- or textual-reasoning. It also obtains state-of-the-art results on a real-world tourism questions dataset, with substantial improvement in answering location questions.

2 RELATED WORK

Our work is related to four broad areas of question answering and information retrieval:

Geographical Information Systems: There is significant prior work on Geographical Information systems where standard IR models are augmented with spatial knowledge [14, 26]. Models have been developed to address challenges in adhoc-retrieval tasks with locative references [15, 24, 29]. However, such models deal primarily with inference problems in toponyms (eg, "Beijing is located in China"), location disambiguation and use of topographical classes (eg, "Union lake is a water-body") etc. Methods for IR involving locative references use three strategies (i) a pipeline of filtering based on spatial information followed by text-based IR (ii) a pipeline of filtering based on text-based IR followed by ranking based on geo-spatial ranking or coverage, and (iii) a weighted or linear combination of two independent rankings [21]. Our work builds on the third strategy by jointly training a model with both geo-spatial and textual components. To the best of our knowledge, joint reasoning over text and geo-spatial data has not been investigated in geographical IR literature.

[2]The word "textual" here does not refer to the fact that questions are textual (which indeed they are). We, instead use "textual reasoning" to imply that, to answer questions, a model has to also reason over a textual corpus (in this case, a corpus of POI reviews)

Geo-Spatial Querying: There has been considerable work in research areas of geo-parsing (toponym discovery and disambiguation) [19], geo-spatial query processing over structured or RDF knowledge bases (KB) [30, 36], geocoding and geo-tagging documents [10, 18, 23] etc. However, such querying methods require KB & task-specific annotations for training and are thus, specialized in application and scope [30].

Numerical Reasoning for Question Answering: Spatial reasoning in our task is effectively a form of numerical reasoning over distances between location-mentions in a question and a candidate entity (POI). Recently introduced tasks such as DROP [12] and QuaRTz [32] require reasoning that includes addition, subtraction, counting, etc. for answering reading comprehension style questions. Other tasks such as MathQA [1] and Math-SAT [17] present high school and SAT-level algebraic word problems.

Models developed for numerical reasoning tasks such as NAQANet [12] and NumNet [27] reason over the explicit mentions of numerical quantities within a question or passage. In contrast, the questions in our task do not explicitly mention geographical coordinates, and also do not contain all the information required for numerical reasoning (since the distances need to be computed with respect to a candidate answer under consideration). Further, in contrast to algebraic word problems and numerical reasoning questions, answers in the POI-recommendation task are also heavily influenced by text-based reasoning on subjective POI-entity reviews.

Points-of-Interest (POI) Recommendation: Existing models for POI recommendation typically rely on the presence of structured data, including geo-spatial coordinates. Queries may be structured or semi-structured and can consist of both spatial as well as textual arguments. Textual arguments are usually associated with the structured attributes or may serve as filters. Approaches include efficient indexing for 'spatial' and 'preference' features along with specialized data-structures as IR-Trees, [7, 22, 33, 39], methods based on Matrix Factorization [38] for user-specific recommendations, click-through logs used for recommendations from search engines [40] etc.

Our work builds on the recently-released POI entity-recommendation QA task [8, 9]. Two approaches have been developed for this task: semantic parsing of unstructured user questions to query a semi-structured knowledge store [8], and an end-to-end trainable neural model operating over a corpus of unstructured reviews to represent POIs [9]. Neither of these approaches explicitly reason on spatial constraints, even though the questions contain them.

Joint Spatio-Textual Reasoning – Challenges: Questions requiring spatio-textual reasoning can be under-specified, ambiguous and also contain subjective spatial constraints – for example, "*near* Times Square", "*driving distance* from the Golden Gate Bridge", "*walking distance* from the Taj Mahal". In addition, they may contain mentions of "distractor locations" i.e, locations that do not require to be reasoned over for answering. Finally, they may contain ambiguous constraints that need to be reasoned over using subjective review (textual) documents for each entity; for example, "a *non-touristy* place with *great* but *cheap* Mexican food".

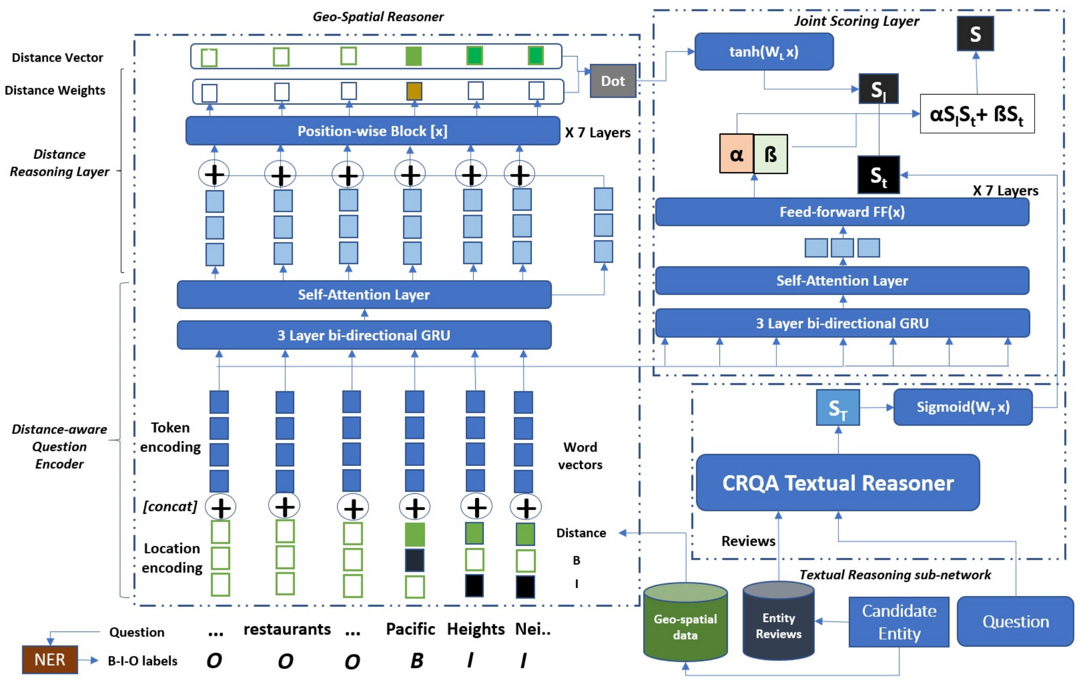

Figure 2: Spatio-Textual reasoning network consisting of (i) Geo-Spatial Reasoner (ii) Textual-Reasoning subnetwork (iii) Joint Scoring Layer

3 SPATIO-TEXTUAL REASONING NETWORK

Answering a question as in Figure 1 requires reasoning on spatial knowledge (for aspects of location mentioned in the question) and POI reviews (for other preferences/constraints specified in the question). Our model design (Figure 2) is motivated by this reasoning - we have a geo-spatial reasoner which uses the question and spatial knowledge (Section 3.1), a textual reasoner based on [9] that uses the question and POI reviews (Section 3.2), and a joint layer (Section 3.3) that combines the reasoning from these two sub-components to do *joint* reasoning.

3.1 Geo-Spatial Reasoner

Our geo-spatial reasoner consists of the following components: (1) **Distance-aware Question Encoder** - to encode questions along with geo-spatial distances between location mentions (in the question) and a candidate entity, (2) **Distance Reasoning layer** - to enable reasoning over geo-spatial distances with respect to the spatial constrains mentioned in the question, (3) **Spatial Relevance Scorer** - to score and rank candidates for spatial-relevance.

Distance-aware Question Encoder: We generate question representations by using embedding representations of their constituent tokens along with embedding representations of their location-mentions. A question token can be represented by traditional word-vector embeddings, or contextual embeddings such as BERT [11]. Each token representation is further appended with a one-hot encoding representing Begin (**B**), Intermediate (**I**) or Other (**O**) labels, indicating the presence of location tokens. The B-I labels help the model recognize a single continuous location-mention. In addition, we concatenate the distance[3] of the candidate entity c, from a location-mention to each token-representation Thus, the question representations are distance-aware and candidate-dependent.

Formally, let the token embedding representations in a question be given by v_i $(v_0 \dots v_i \dots v_{m-1})$,where m is the length of a question. Let the distance between the kth location-mention lm_k and c be denoted by d_k. Further, let $\phi(lm_k)$ be a function that returns the set of position indices occupied by location mention lm_k, i.e. it returns the set of position indices of question tokens that have been assigned the B or I label from the B-I encoding for location mention lm_k, $(\phi(lm_k) \subset \{0, \dots, m-1\})$. We create an m-dimensional distance vector $\mathbf{d'}$ where each element d'_i of the vector is given by:

$$d'_i = \begin{cases} d_k & \text{if i} \in \phi(lm_k) \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Let the one-hot vector (two dimensional) of the B-I labels for the ith position be g_i. The input question embedding t_i, $(t_0 \dots t_i \dots t_{m-1})$ is then given by:

$$t_i = concat[v_i, d'_i, g_i] \tag{2}$$

We encode the question using a bi-directional GRU [6] which results in output states $\mathbf{q}$ $(q_0 \dots q_i \dots q_{m-1})$.

Distance-Reasoning Layer (DRL): We first used a series of down-projecting feed-forward layers applied to the output state of the GRU, to generate the final score for each candidate, but we found this was not effective (Section 4.1.2). We therefore include a component designed for distance-reasoning referred to as the 'Distance Reasoning Layer' which uses the representations generated by the distance-aware question encoder.

A model could score candidate-entities for relevance if, for each location mentioned in the question, it is able to (i) learn whether a location-mention needs to be considered for answering, and (ii)

[3]Manhattan Distance

learn *how* a location-mention needs to be used for answering. Our design of the DRL is motivated by this insight – it learns a function which, for *each* location-mention lm_k in the question, outputs a *distance-weight* w_k. Here, w_k captures the contribution of the spatial-distance between lm_k and the candidate entity c, under the constraints mentioned in the question. For instance, a question may include location-mentions that could be involved in simple 'near' or 'far' constraints or other complex constraints such as "within driving distance" or "within walking distance" etc. The DRL layer uses the distance-aware question encoding to understand the nature of the constraint being expressed, as well as, figure out how to compute distance-reasoning weights to express those constraints.

Let the output states of the question encoder be given by $q_0 \,..q_i..\, q_{m-1}$, where m is the length of the question. To compute distance-weights, we use a series of position-wise feed-forward blocks [34] that consist of a linear layer with ReLU activation applied at each output position of the Question Encoder:

$$q_i^l = Block_l(q_i^{l-1}) = max(0, A_l q_i^{l-1} + b_l) \tag{3}$$

where q_i^l is the output of the Block layer at layer l, A_l is a weight matrix and b_l the bias term for each layer l.

The initial block input uses the output state of the GRU (q_i) concatenated with the final hidden state (q_{m-1}). Thus, the output q_i^1 from the application of the first block layer, corresponding position i in the input is given by:

$$q_i^1 = Block_1(concat[q_i, q_{m-1}]) \tag{4}$$

The blocks apply the same linear transformations at each position but we vary the parameters across layers (see appendix). The final layer gives us a single dimension output for each position resulting in an m-dimensional vector $\mathbf{r}$ ($r_0...r_i..r_{m-1}$).

Let $\mathbf{B}$ be an m-dimensional one hot-vector based on the position indices that have been assigned only the B label[4] from the *B-I* encoding used in the input layer. The distance-weight vector $\mathbf{w}$ for a question is given by:

$$\mathbf{w} = tanh(r \odot B) \tag{5}$$

We use the distance-weights for scoring, as described below.

Spatial Relevance Scorer: The final score S_L of a candidate c is given by:

$$S_L = \mathbf{w}\dot{\mathbf{d}}' \tag{6}$$

Note that since we concatenate the distance values along with token embeddings while encoding locations as part of the Question Encoder (Equation 2), it helps learn distance weights w which are dependent on the distance value as well as the semantic information present in the question. Thus, the spatial relevance score is not just a simple linear combination of distances and makes the model representationally more powerful (see experiments in Section 4.2). We refer to the Geo-Spatial Reasoner as SPNet for brevity in the rest of the paper.

[4]An element of $\mathbf{B}$ is 1 whenever it corresponds to a position index indicating the *start* of a location mention in a question.

3.2 Textual-Reasoning Sub-network

We use the CrQA [9] model as our textual-reasoning sub-network. It consists of a Siamese-Encoder [20] that uses question representations to attend over entity-review sentences and generate question-aware entity-embeddings. These entity embeddings are combined with question representations to generate an overall relevance score. For scalability, instead of using full review documents, the model uses a set of representative sentences from reviews after clustering them in USE-embedding space [3]. We follow Contractor et al. [9] and use k-means to cluster sentences in USE embedding space. We set k=10, and select 10 sentences per cluster, thus creating a ≤ 100-sentence document to represent an entity. In order to build a model that is capable of joint spatio-textual reasoning, our model learns question-specific combination weights that combine textual and spatial-reasoning scores.

3.3 Joint Scoring Layer

Let the score generated by the textual-reasoner be S_T and let the score generated by the spatial-reasoner be S_L. Let the rescaling weights for S_T and S_L be w_T and w_L respectively. Then, the overall score S is given by:

$$\alpha.\sigma(w_T S_T).\tanh(w_L S_L) + \beta.\sigma(w_T S_T),$$

where σ is the Sigmoid function and α,β are combination weights. The weights are computed by returning a two dimensional output (corresponding to each weight), after a series of feed-forward operations on the self-attended representation [5], of the question using the outputs of a Question Encoder with the same architecture as in SPNet (see appendix for hyperparameters). Note that the first term of scoring equation uses S_L as a *selector* – for questions where there are no locations mentioned, the spatial score of a question with no location-mentions will be 0 (due to the equation for $\mathbf{w}$). This lets the model rely only on textual scores for these cases.

Training: We train the joint model using max-margin loss, teaching the network to score correct-answer entities higher than negative samples.

4 EXPERIMENTS

We first present a detailed study of the spatial-reasoner using a simple artificially generated toy-dataset. This allows us to probe and study different aspects of spatial-reasoning in the absence of textual reasoning. We then present our experiments with the joint spatio-textual model using a real-world POI-recommendation QA dataset (Section 4.2)

4.1 Detailed Study: Geo-Spatial Reasoner

We conduct this study on a simple synthetic-dataset generated using linguistically diverse templates specifying spatial constraints and location names chosen at random from a list of 200, 000 entities across 50 cities.

4.1.1 Synthetic-Dataset. **Template Classes:** We create templates that can be broadly divided into three types of proximity queries based, on whether the correct answer entity is expected to be: (1) close to one or more locations (mentioned in the question), (2) far from one or more locations, (3) close to some and far from others (combination). We create different templates for each category

Table 1: Results of SPNet on the synthetic spatial-questions dataset (t-test p-value $< 10^{-33}$ for Hits@3)

	Close to set X			Far from set X			Combination			Aggregate		
Models	**Hits@3**	**MRR**	$Dist_g$	**Hits@3**	**MRR**	$Dist_g$	**Hits@3**	**MRR**	$Dist_g$	**Hits@3**	**MRR**	$Dist_g$
SPNet w/o DRL	62.60	0.608	2.88	89.00	0.858	15.24	23.40	0.229	9.72	58.33	0.565	9.28
SPNet	90.20	0.873	0.860	**98.00**	0.975	13.88	52.80	0.486	3.90	80.33	0.778	6.21
BERT SPNet w/o DRL	63.60	0.616	3.68	90.80	0.881	15.32	26.80	0.242	12.96	60.40	0.579	10.65
BERT SPNet	**91.40**	**0.896**	**0.78**	97.80	**0.978**	**13.87**	**59.20**	**0.551**	**3.02**	**82.80**	**0.808**	**5.89**

Question 1: "Suggestions for a place of tourist interest close to Hilal Park and Mahatta Palace Museum"
Question 2: "Hey I will be staying at Pinati. Please suggest a coffee shop far from the Cellar Bar."
Question 3: "Please propose an eatery close to Udipi Palace but not in the neighbourhood of Nico. Thanks!"

Figure 3: Sample questions from the synthetic dataset. The dataset has questions from three categories: (1) close to set X, (2) far from set X (3) Combination.

with linguistics variations. Figure 3 shows a sample question from each category. See appendix for more details, including the list of templates.

Use of distractor-locations: In order to make the task more reflective of real-world challenges we also randomly insert a *distractor* sentence that contains a location reference which does not need to be reasoned over (e.g the location "Pinati" in Question 2 in Figure 3).

Gold-entity generation: The gold answer entity is uniquely[5] determined for each question based on its template. For example, consider a template T, "*I am staying at $A! Please suggest a hotel close to $B but far from $K.*" The score of a candidate entity X is given by $dist_T(X) = -(dist(X,B) - dist(X,K))$ (distances from B needs to be reduced, while distance from K needs to be higher). A is a distractor. The candidate with the $max(dist_T(X))$ in the universe is chosen as the gold-answer entity for that question. We use the geo-coordinates of locations to compute the distance.

Dataset Statistics: The train, dev and test sets consist of 6000, 1500 and 1500 questions respectively generated using 48 different templates, split equally across all 3 template categories. Each question consists of location-names from only one city and thus the candidate search space for that question is restricted to that city. The average search space for each question is 1250, varying between 10-16200 across cities. The dataset includes questions containing distractor-locations (52.33% of dataset) distributed evenly across all template classes.

4.1.2 Results. We study SPNet using the synthetic dataset to answer the following questions: (1) What is the model performance across template classes? (2) How does the network compare with baseline models that do not use the DRL? (3) How well does the model deal with distractor-locations, i.e locations not relevant for the scoring task? For all experiments in this section we use perfectly tagged location-mentions.

[5]In general, a question may have many answers. The goal of the synthetic dataset was only to test that the model is capable of spatial reasoning. So, for simplicity, we defined the task of predicting the "best" POI entity.

Table 2: Performance of spatial-reasoning networks degrades in the presence of location-distractor sentences.

	Without Distractors		With Distractors	
Models	**Hits@3**	**MRR**	**Hits@3**	**MRR**
SPNet	82.58	0.800	78.30	0.758
BERT SPNet	**84.13**	**0.820**	**81.60**	**0.797**

Metrics: We study the performance of models using $Hits@N$ (N=3,5, 30)[6] which requires that any one of the top-N answers be correct, Mean Reciprocal Rank (*MRR*) and the average distance of the top-3 ranked answers from the gold-entity $Dist_g$. $Dist_g$ is helpful in quantifying the spatial goodness of the returned answers (lower is better).

We use the following models in our experiments: (i) SPNet (ii) SPNet without DRL (iii) BERT-SPNet (uses contextual BERT embeddings instead of word vectors while encoding questions) (iv) BERT-SPNet without DRL. Models without DRL use the final hidden states of the Question Encoder and a series of down-projecting feed-forward layers to generate the final score.

Performance across template classes: As can be seen in Table 1, all models perform the worst on the template class that contains a combination of both 'close-to' and 'far' constraints. Models based on SPNet perform exceeding well on the 'Far' templates because the difference between the $dist_T(X)$ scores of the best and the second best candidate is almost always large enough for every model to easily separate them.

Importance of Distance-Reasoning Layer: As can be seen in Table 1 the performance of each configuration (with and without BERT) suffers a serious degradation in the absence of the DRL. Recall, that all models have access to spatial knowledge in their input layer via the question encoding. This indicates that the DRL is an important component required for reasoning on spatial constraints. To further assess whether our model is able to do distance reasoning, we computed the correlation between ranking-by-distances (appropriate ranking order for each template-class) and SPNet's ranking on the synthetic-dataset. We found the rank correlation to be a high 0.97 ($p < 0.002$) suggesting that the model is able to use physical distance to compute the best answer.

Effect of distractor-locations: We report results on two splits of the test set: Questions with and without distractor-locations. We report the aggregate performance over all template classes due to space constraints. As can be seen in Table 2, models suffer a degradation of performance in the presence of distractor-locations. We hypothesize that this is because the reasoning task becomes harder; models now need to also account for location-mentions that do not need to be reasoned over.

[6]We report results with N=3 in the main paper. Please see appendix for full results.

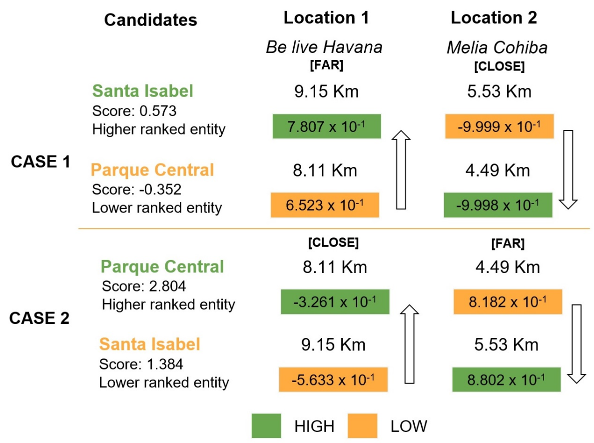

Figure 4: Probing study of the Distance Reasoning Layer (DRL) using the question: *"I came from Tropicoco today. Any nice ideas for a coffee shop [*`far from/close to`*] 'Be Live Havana' but [*`close to/far from`*] 'Melia Cohiba'?"*. The coloured boxes indicate the relative magnitude of weights assigned; each candidate entity assigns a *higher* weight (column-wise comparison), as compared to the other candidate, on the distance property it is most likely to benefit from, with respect to the spatial-constraint

Probing Study: We conduct a probing study (Figure 4) on SPNet to get some insights into the reasoning process employed by the trained network. We use a question that has both 'near' and 'far' constraints (case 1) and then interchange the constraints (case 2). In both the cases we study the corresponding distance-weights assigned to the location-mentions with respect to two candidates "Santa Isabel" and "Parque Central". Consider the first case; as can be seen, each candidate entity assigns a *higher* weight (column-wise comparison) as compared to the other candidate, on the distance property it is most likely to benefit from, with respect to the spatial-constraint. For example, when the spatial-constraint requires an answer to be *close* to "Melia Cohiba", the candidate "Parque Central" assigns a higher weight to this location as compared to candidate "Santa Isabel", since "Parque Central" has a *smaller* distance value to this location. On the other hand, with respect to the "*far*" constraint, candidate "Santa Isabel" has a *larger* distance value from "Be live Havana" as compared to candidate "Parque Central", thus assigning a *higher* distance weight for this location-mention.

When we interchange the constraints (Case 2) we see the same pattern and the comparative weight trends (at each location-mention) invert due to inversion of spatial-constraints. This suggests, that DRL is learning to transform the inputs and generate weights based on the spatial constraint at hand.

Effect of Candidate Space Size: We analyzed the errors made by the SPNet model and we find that nearly 40% of the errors were made in questions that have large (> 1000) candidate spaces. Approximately 25% of the test-set contains questions with large candidate spaces.

Effect of the No. of Location-mentions: The complexity of the spatial-reasoning task increases as the number of location-mentions (including distractor-locations) in the question increase. We find that SPNet makes no errors when spatial-reasoning involves only 1 location-mention but, nearly 57% of the errors are made in questions with 3 location-mentions (See appendix).

Table 3: Dataset statistics: Questions with and without location-mentions across train, dev & test sets from [9].

	Location		Non-location	
Dataset	Questions	QA pairs	Questions	QA pairs
Train	9,617	21,396	10,342	22,150
Dev	1,065	2,209	1,054	1,987
Test	1,086	2,198	1,087	2,144

4.2 Spatio-Textual Reasoning Network

For the joint model, we investigate the following research questions: (i) Does joint spatio-textual ranking result in improved performance over a model with only spatial-reasoning or only textual-reasoning? (ii) How do pipelined baseline models that use spatial re-ranking perform on the task? (iii) Does distance-aware question encoding help in spatio-textual reasoning? (iv) Is the spatio-textual reasoning model more robust to distractor-locations as compared to baselines? (v) What kind of errors does the model make?

Dataset: We use the recently released data set[7] on Tourism Questions [9] that consists of over 47,000 real-world POI question-answer pairs along with a universe of nearly 200,000 candidate POIs; questions are long and complex, as presented in Figure 1, while the recommendations (answers) are represented by an ID corresponding to each POI. Each POI comes with a collection of reviews and meta-data that includes its geo-coordinates. The training set contains nearly $38,000$ QA-pairs and about $4,200$ QA-pairs each in the validation and test sets. The average candidate space for each question is $5,300$.

Task Challenges: The task presents novel challenges of reasoning and scale; the nature of entity reviews (eg. inference on subjective language, sarcasm etc) makes methods such as BM25 [28], that are often used to prune the search space quickly in large scale QA tasks [4, 13], ineffective. Thus, even simple BERT-based architectures or popular models such as BiDAF [31] do not scale for the answering task in this dataset [9].Thus, we use the non-BERT based SPNet subnetwork in the rest of the QA experiments[8].

Evaluation Challenges: It is infeasible to construct a dataset of POI recommendation QA pairs, which has an exhaustively labeled answer-set for each question, since the candidate space is very large. Hence, this dataset suffers from the problem of false negatives, and *Hits@N* metrics under-report system performance. Still, they are shown to be correlated with human relevance judgments [9]. We therefore, use these metrics for all experiments, but additionally present a small human-study on the end-task, verifying the robustness of our results.

Location Tagging in Questions: In order to get mentions of locations in questions, we manually label a set of 425 questions from the training set for location mentions. We then use a BERT-based sequence tagger[9] trained on this set to label locations. The tagger has a macro-$F1$ of 88.03. This tagger tags all location mentions in

[7] https://github.com/dair-iitd/TourismQA
[8] CrQA is also not based on BERT due to this reason.
[9] https://github.com/dair-iitd/LocationTagger

Table 4: Comparison of the joint Spatio-Textual model with baselines on questions that have location mentions (t-test p-value< 0.009)

	Location Questions				
Models	**Hits@3**	**Hits@5**	**Hits@30**	**MRR**	**Dist$_g$**
SD	2.49	3.41	14.29	0.029	3.07
SPNet	1.47	2.11	8.47	0.019	2.97
CRQA	14.83	21.27	50.65	0.143	3.41
CRQA→SD	13.73	19.26	50.65	0.125	**2.23**
CRQA→SPNet	10.13	15.65	50.64	0.104	2.47
Spatio-textual CRQA	**18.32**	**25.69**	**56.17**	**0.168**	2.62

a question without considering their utility for spatial-reasoning. Thus, it is possible that a question may contain **only** distractor-locations, i.e., locations-mentions that do not need to be reasoned over the answering task.

Once the location-mentions are tagged, we remove the punctuations and stopwords from the tagged-location span. We then query the Bing Maps Location API[10] using the location-mention, along with the city (known from question meta-data), to get the geo-tags. To reduce noise in geo-tagging, we ignore the location-mention if the remaining text has a length of less than 4 characters or is identified as a popular acronym, continent, country, city or state (lists from Wikipedia). We further reduce noise by ignoring a location mention: (1) if no results were found from Bing, or (2) if the geo-tag is beyond 40km from the city center. We found the location-mention geo-tagging precision on a small set of 83 location-mentions to be 96%.

We label all questions in the full dataset using this tagger, resulting in approximately 49.54% of the QA pairs containing at least one location-mention (see Table 3). In all our experiments, we use the Manhattan distance as our distance value, because it is generally closer to real-world driving/walking distance within a city, as opposed to straight-line distance.

4.2.1 Baselines. Apart from the textual-reasoning model CrQA we also use the following baselines in our experiments:

Sort-by-distance (SD): Given a set of tagged-locations in a question and their geo-coordinates, rank candidate entities by the *minimum* distance from the set of tagged locations.

SPNet : Use only the spatial-reasoning network for ranking candidate entities using their geo-coordinates. No textual-reasoning performed.

CrQA → SD: Rank candidates using CrQA and then re-rank the top-30 answers using SD.

CrQA → SPNet : Rank candidates using CrQA and then re-rank the top-30 answers using SPNet.

Training: We pretrain SPNet on this dataset by allowing entities within a radius of 100m from the actual gold-entity to be considered as gold (only for pretraining). To train the joint network we initialize model parameters learnt from component-wise pretraining of both SPNet as well as CrQA.

4.2.2 Results. We present our experiments on two slices of the test-set – questions with tagged location-mentions (called *Location-Questions*) and those without any location mentions (*Non-Location Questions*). As can be seen in Table 4 sorting-by-distance (SD) performs very poorly indicating that simple methods for ranking based on entity-distance do not work for such questions. Further, the poor performance of SPNet also indicates that the task cannot be solved just by reasoning on location data.

In addition, pipelined re-ranking using SD or SPNet over the textual reasoning model decreases the average distance ($Dist_g$) from the gold-entity but does not result in improved performance in terms of answering (Hits@N) indicating the need for spatio-textual reasoning. Finally, from Tables 4 & 5 we note that the spatio-textual model performs better than its textual counterpart on the Location-Questions subset, while continuing to perform well on questions without location mentions.

Effect of distance-aware question encoding: In order to demonstrate the importance of distance-aware question encoding, we present an experiment where we remove the distance values from the input encoding. Thus, Equation 2 changes to $t_i = concat[v_i, g_i]$. As Table 5 shows, the performance of the Spatio-Textual CrQA model in the absence of distance-aware encoding drops (last row), but it still performs better than the text-only CrQA model (first row). This indicates that the distance-aware question encoding helps learn better distance weights for spatio-textual reasoning.

Effect of distractor-locations: As mentioned earlier, we use a location-tagger that is oblivious to the reasoning task, to tag locations in the dataset. We manually create a small set of 200 questions, randomly selected from the test-set, but ensuring that half of it contains *at least* one non-distractor location mentioned in the question while the other half contains questions with *only* distractor-locations.

As can be seen from Table 6, all models including the spatio-textual model deteriorate in performance if a question only contains distractors; the spatio-textual model however, suffers a less significant drop in performance.

Qualitative Study: We randomly selected 150 QA pairs with location-mentions from the test-set, to conduct a qualitative error analysis of Spatio-textual CrQA (Table 7). We find that nearly 37% of the errors can be traced to the textual-reasoner, 22% of the errors were due to a 'near' constraint not being satisfied, while about 13% of the errors were due to the model reasoning on distractor-locations. Lastly 8% of the errors were due to errors made by the location-tagger and incorrect geo-spatial data.

Effect of Candidate Search Space: Past work [9] has improved overall task performance by employing a neural IR method to reduce the search space [25], and then using the CrQA textual-reasoner to re-rank only the top 30 selected candidates (pipeline referred to as CsrQA). In line with their work, we create a spatio-textual counterpart to CsrQA, by using spatio-textual reasoning in re-rank step. We find that this final model results in a 1 pt (Hits@3) improvement overall (see Table 16), and a 1.5 pt improvement on location questions (Hits@3), establishing a new state of the art on the task. We note that, because the IR selector is incapable of spatial-reasoning, it possibly reduces the gains made by the spatio-textual re-ranking. An interesting direction of future work could be to augment general purpose neural IR methods with such spatial-reasoning.

Effect of False Negatives: To supplement the automatic evaluation, we additionally conducted a blind human-study using the

[10]https://bit.ly/36Vazwo

Table 5: Comparison of Spatio-Textual CrQA (with and without (w/o) distance-aware question encoding) and CrQA (t-test p-value < 0.03 for $Hits@3$)

Models	Location Questions					Non-location Questions				Full Set			
	Hits@3	**Hits@5**	**Hits@30**	**MRR**	$Dist_g$	**Hits@3**	**Hits@5**	**Hits@30**	**MRR**	**Hits@3**	**Hits@5**	**Hits@30**	**MRR**
Spatio-Textual CRQA	**18.32**	**25.69**	**56.17**	**0.168**	**2.62**	**20.42**	**26.77**	**56.49**	**0.18**	**19.37**	**26.23**	**56.33**	**0.175**
Spatio-textual CrQA (w/o distance-aware QE)	16.85	23.39	53.04	0.159	2.84	20.06	26.86	56.49	0.185	18.45	25.13	54.76	0.172

Table 6: Experiments on two subsets from the test-set: (i) Questions requiring Spatial-reasoning (ii) Questions with distractor-locations only.

	Questions requiring Spatial-reasoning				
Models	**Hits@3**	**Hits@5**	**Hits@30**	**MRR**	$Dist_g$
SD	5.00	7.00	22.00	0.053	2.10
SPNet	1.00	1.00	8.00	0.013	2.64
CRQA	15.00	17.00	51.00	0.132	3.53
CRQA→SD	15.00	22.00	51.00	0.142	**1.963**
CRQA→SPNet	16.00	23.00	51.00	0.134	2.41
Spatio-textual CRQA	**22.00**	**28.00**	**54.00**	**0.182**	2.62
	Questions with distractor-locations only				
SD	2.00	3.00	17.00	0.025	4.12
SPNet	1.00	2.00	9.00	0.016	4.14
CRQA	19.00	26.00	51.00	0.162	3.62
CRQA→SD	13.00	17.00	51.005	0.108	3.26
CRQA→SPNet	13.00	17.00	51.00	0.113	**3.24**
Spatio-textual CRQA	**20.00**	**28.00**	**53.00**	**0.187**	3.50

Table 7: Spatio-Textual CrQA: Classification of Errors

Error Type	**Percentage**
Textual Reasoning Error	37.9%
Far from the required location	22.3%
Influenced by Distractor	12.6 %
Not in requested Neighbourhood	10.7 %
Location Tagger Error	5.8 %
RepeatedLocation Names	4.9 %
Error in Geo-Spatial Data	2.9 %
Invalid Question	2.9 %

Table 8: Comparison with current state-of-the-art CsrQA on (i) Location Questions (ii) All data

	Location Questions				
Models	**Hits@3**	**Hits@5**	**Hits@30**	**MRR**	$Dist_g$
CsrQA	19.89	26.43	**51.47**	0.168	2.70
Spatio-textual CSRQA	**21.36**	**28.36**	**51.47**	**0.183**	2.27
	All Questions				
CsrQA	21.45	28.21	**52.65**	0.186	2.47
Spatio-textual CSRQA	**22.41**	**28.99**	**52.65**	**0.193**	**2.32**

Table 9: Hits@3 results on a blind-human study using 100 **randomly selected questions from the test-set**

	Automated evaluation		**Human evaluation**	
	Location	**Non-location**	**Location**	**Non-location**
CsrQA	28.00	36.00	64.00	70.00
Spatio-textual CSRQA	32.00	32.00	84.00	72.00

top-ranked CsrQA and spatio-textual CsrQA models on another subset of 100 questions from the test-set. Two human evaluators (κ=0.81) were presented the top-3 answers from both models in random order and were asked to mark each answer for relevance. As Table 9 shows, the manual annotation resulted in Hits@3 for CsrQA and spatio-textual CsrQA at a much higher, 67% and 78% respectively. On the subset of location questions, the accuracy numbers are 64% and 84%. This underscores the value of joint spatio-textual reasoning for the task, and indicates a substantial improvement in the overall QA performance.

5 CONCLUSION

Our paper presents the first joint spatio-textual QA model that combines spatial and textual reasoning. Experiments on an artificially constructed (spatial-only) QA-dataset show that our spatial reasoner effectively trains to satisfy spatial constraints. We also presented detailed experiments on the recently released POI recommendation task for tourism questions. Comparing against textual only and spatial only QA models, the joint model obtains significant improvements. Our final model establishes a new state of the art on the task. In future work, we would like to also support reasoning on questions that require directional or topographical inference (eg.“north of X”, “on the river beach”). In addition, our work is one instance of the general problem of joint reasoning over a knowledge base and text. Even though we only considered spatio-textual reasoning, we believe that our model has parallels in other domains - for example, questions seeking product recommendations such as “*a laptop with a responsive keyboard and* 14-*inch screen*” will require a system to reason over screen size using a database of product properties along with textual reasoning about, whether or not, the keyboard is responsive (likely found in user reviews). Resources from our work are available at: https://ibm.biz/SpatioTextualQA

ACKNOWLEDGMENTS

We would like to thank Krunal Shah, Gaurav Pandey, Biswesh Mohapatra, Azalenah Shah for their helpful suggestions to improve early versions of this paper. We would like to thank Jochen Leidner and Bruno Martins for their help in contextualizing our work. We would also like to acknowledge the IBM Research India PhD program that enables the first author to pursue the PhD at IIT Delhi. This work was supported by an IBM AI Horizons Network (AIHN) grant, IBM SUR awards, Visvesvaraya faculty awards by Govt. of India to both Mausam and Parag, as well as grants by Google, Bloomberg and 1MG to Mausam.

REFERENCES

[1] Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards Interpretable Math Word Problem Solving with Operation-Based Formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 2357–2367. https://doi.org/10.18653/v1/N19-1245

[2] Bin Bi, Chen Wu, Ming Yan, Wei Wang, Jiangnan Xia, and Chenliang Li. 2019. Incorporating External Knowledge into Machine Reading for Generative Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. 2521–2530. https://doi.org/10.18653/v1/D19-1255

[3] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder. *CoRR* abs/1803.11175 (2018). arXiv:1803.11175 http://arxiv.org/abs/1803.11175

[4] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Association for Computational Linguistics (ACL)*.

[5] Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long Short-Term Memory-Networks for Machine Reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 551–561. https://doi.org/10.18653/v1/D16-1053

[6] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1724–1734. https://doi.org/10.3115/v1/D14-1179

[7] Gao Cong, Christian S. Jensen, and Dingming Wu. 2009. Efficient Retrieval of the Top-k Most Relevant Spatial Web Objects. *Proc. VLDB Endow.* 2, 1 (Aug. 2009), 337–348. https://doi.org/10.14778/1687627.1687666

[8] Danish Contractor, Barun Patra, Mausam, and Parag Singla. 2021. Constrained BERT BiLSTM CRF for understanding multi-sentence entity-seeking questions. *Natural Language Engineering* 27, 1 (2021), 65–87. https://doi.org/10.1017/S1351324920000017

[9] Danish Contractor, Krunal Shah, Aditi Partap, Mausam, and Parag Singla. 2019. Large Scale Question Answering using Tourism Data. *CoRR* abs/1909.03527 (2019). arXiv:1909.03527 http://arxiv.org/abs/1909.03527

[10] Gaétan De Rassenfosse, Jan Kozak, and Florian Seliger. 2019. Geocoding of worldwide patent data. *Scientific data* 6, 1 (2019), 1–15.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[12] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In *Proc. of NAACL*.

[13] Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Güney, Volkan Cirik, and Kyunghyun Cho. 2017. SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine. *CoRR* abs/1704.05179 (2017). arXiv:1704.05179 http://arxiv.org/abs/1704.05179

[14] Daniel Ferrés Domènech. 2017. Knowledge-based and data-driven approaches for geographical information access. (2017).

[15] Fredric Gey, Ray Larson, Mark Sanderson, Kerstin Bischoff, Thomas Mandl, Christa Womser-Hacker, Diana Santos, Paulo Rocha, Giorgio M Di Nunzio, and Nicola Ferro. 2006. GeoCLEF 2006: the CLEF 2006 cross-language geographic information retrieval track overview. In *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 852–876.

[16] Yangyang Guo, Zhiyong Cheng, Liqiang Nie, Xin-Shun Xu, and Mohan Kankanhalli. 2018. Multi-modal preference modeling for product search. In *Proceedings of the 26th ACM international conference on Multimedia*. 1865–1873.

[17] Mark Hopkins, Ronan Le Bras, Cristian Petrescu-Prahova, Gabriel Stanovsky, Hannaneh Hajishirzi, and Rik Koncel-Kedziorski. 2019. SemEval-2019 Task 10: Math Question Answering. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 893–899. https://doi.org/10.18653/v1/S19-2153

[18] Binxuan Huang and Kathleen Carley. 2019. A Hierarchical Location Prediction Neural Network for Twitter User Geolocation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 4732–4742. https://doi.org/10.18653/v1/D19-1480

[19] Tannon Kew, Anastassia Shaitarova, Isabel Meraner, Janis Goldzycher, Simon Clematide, and Martin Volk. 2019. Geotagging a Diachronic Corpus of Alpine Texts: Comparing Distinct Approaches to Toponym Recognition. In *Proceedings of the Workshop on Language Technology for Digital Historical Archives*. 11–18.

[20] Tuan Manh Lai, Trung Bui, and Sheng Li. 2018. A Review on Deep Learning Techniques Applied to Answer Selection. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*. 2132–2144. https://aclanthology.info/papers/C18-1181/c18-1181

[21] Jochen L. Leidner, Bruno Martins, Katherine McDonough, and Ross S. Purves. 2020. Text Meets Space: Geographic Content Extraction, Resolution and Information Retrieval. In *Advances in Information Retrieval*, Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (Eds.). Springer International Publishing, Cham, 669–673.

[22] Miao Li, Lisi Chen, Gao Cong, Yu Gu, and Ge Yu. 2016. Efficient Processing of Location-Aware Group Preference Queries. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (Indianapolis, Indiana, USA) *(CIKM '16)*. Association for Computing Machinery, New York, NY, USA, 559–568. https://doi.org/10.1145/2983323.2983757

[23] Kwan Hui Lim, Shanika Karunasekera, Aaron Harwood, and Yasmeen George. 2019. Geotagging Tweets to Landmarks Using Convolutional Neural Networks with Text and Posting Time. In *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 61–62. https://doi.org/10.1145/3308557.3308691

[24] Thomas Mandl, Paula Carvalho, Giorgio Maria Di Nunzio, Fredric Gey, Ray R Larson, Diana Santos, and Christa Womser-Hacker. 2008. GeoCLEF 2008: the CLEF 2008 cross-language geographic information retrieval track overview. In *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 808–821.

[25] Bhaskar Mitra and Nick Craswell. 2019. An Updated Duet Model for Passage Re-ranking. *arXiv preprint arXiv:1903.07666* (2019).

[26] R. S. Purves, P. Clough, C. B. Jones, M. H. Hall, and V. Murdock. 2018. *Geographic Information Retrieval: Progress and Challenges in Spatial Search of Text*.

[27] Qiu Ran, Yankai Lin, Peng Li, Jie Zhou, and Zhiyuan Liu. 2019. NumNet: Machine Reading Comprehension with Numerical Reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 2474–2484. https://doi.org/10.18653/v1/D19-1251

[28] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 3, 4 (April 2009), 333–389. https://doi.org/10.1561/1500000019

[29] Diana Santos and Luís Miguel Cabral. 2009. GikiCLEF: Expectations and lessons learned. In *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 212–222.

[30] Simon Scheider, Enkhbold Nyamsuren, Han Kruiger, and Haiqi Xu. 2020. Geo-analytical question-answering with GIS. *International Journal of Digital Earth* 0, 0 (2020), 1–14. https://doi.org/10.1080/17538947.2020.1738568 arXiv:https://doi.org/10.1080/17538947.2020.1738568

[31] Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional Attention Flow for Machine Comprehension. *CoRR* abs/1611.01603 (2016). arXiv:1611.01603 http://arxiv.org/abs/1611.01603

[32] Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. 2019. QuaRTz: An Open-Domain Dataset of Qualitative Relationship Questions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. 5940–5945. https://doi.org/10.18653/v1/D19-1608

[33] George Tsatsanifos and Akrivi Vlachou. 2015. On Processing Top-k Spatio-Textual Preference Queries. In *Proceedings of the 18th International Conference on Extending Database Technology, EDBT 2015, Brussels, Belgium, March 23-27, 2015*. 433–444. https://doi.org/10.5441/002/edbt.2015.38

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[35] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6439–6448.

[36] Dimitri Vorona, Andreas Kipf, Thomas Neumann, and Alfons Kemper. 2019. DeepSPACE: Approximate Geospatial Query Processing with Deep Learning. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 500–503.

[37] Jiangnan Xia, Chen Wu, and Ming Yan. 2019. Incorporating Relation Knowledge into Commonsense Reading Comprehension with Multi-Task Learning. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (Beijing, China) *(CIKM '19)*. Association for Computing Machinery, New York, NY, USA, 2393–2396. https://doi.org/10.1145/3357384.3358165

[38] M. L. Yiu, X. Dai, N. Mamoulis, and M. Vaitis. 2007. Top-k Spatial Preference Queries. In *2007 IEEE 23rd International Conference on Data Engineering*. 1076–1085.

[39] C. Zhang, Y. Zhang, W. Zhang, and X. Lin. 2016. Inverted Linear Quadtree: Efficient Top K Spatial Keyword Search. *IEEE Transactions on Knowledge and Data Engineering* 28, 7 (2016), 1706–1721.

[40] Ji Zhao, Meiyu Yu, Huan Chen, Boning Li, Lingyu Zhang, Qi Song, Li Ma, Hua Chai, and Jieping Ye. 2019. POI Semantic Model with a Deep Convolutional Structure. *CoRR* abs/1903.07279 (2019). arXiv:1903.07279 http://arxiv.org/abs/1903.07279

Table 10: Templates used for generating the synthetic dataset

Id	Description
1	Do you have any recommendations of *ENTITY* near the *LOCATION*?
2	Does anyone have ideas on *ENTITY* close to *LOCATION*? Thank you!
3	Hello! Could anyone please suggest *ENTITY* in the neighborhood of *LOCATION*?
4	Good Morning! Can someone please propose *ENTITY* not very far from *LOCATION*?
5	Suggestions for *ENTITY* close to both *LOCATION* and *LOCATION*?
6	Some good ideas of *ENTITY* between *LOCATION* and *LOCATION*? Thanks much!
7	Please advise *ENTITY* close to *LOCATION* and not very far off the *LOCATION.*
8	Any ideas for *ENTITY* near *LOCATION* and also close to *LOCATION* would be welcomed?
9	I once lived around *LOCATION.* Does anyone have ideas of *ENTITY* close to the *LOCATION*? Thanks!
10	Any nice suggestions of *ENTITY* near the *LOCATION*? I will be going to *LOCATION* the next day.
11	I just came from *LOCATION.* Someone, please recommend *ENTITY* in the neighborhood of *LOCATION.*
12	Could anyone propose *ENTITY* not far from the *LOCATION*? I need to leave for *LOCATION* urgently.
13	We came from *LOCATION* this morning. Suggestions for *ENTITY* close to both *LOCATION* and *LOCATION*?
14	Any ideas of *ENTITY* between *LOCATION* and *LOCATION*? I would be going to *LOCATION.* Thanks.
15	We might be staying around *LOCATION.* Please advise *ENTITY* close to *LOCATION* and not far from *LOCATION.*
16	Could anyone suggest ideas for *ENTITY* close to *LOCATION* and around *LOCATION*? We could be going to *LOCATION* soon.
17	Any suggestions for *ENTITY* quite far from the *LOCATION*? Thank you very much!
18	Somebody please suggest *ENTITY* cut off from *LOCATION.* Have a good day!
19	Does anyone have suggestions for *ENTITY* away from *LOCATION*? Thanks a lot!
20	Good Afternoon! Any proposals for *ENTITY* not very close to the *LOCATION*?
21	Suggestions on *ENTITY* far from both *LOCATION* and *LOCATION*? Thank!
22	Hi! Any idea of *ENTITY* far away from *LOCATION* and *LOCATION*?
23	Could anyone please propose *ENTITY* not close to *LOCATION* and also far from *LOCATION*?
24	Does anyone have any suggestions for *ENTITY* far from *LOCATION* and not around *LOCATION*?
25	Hey! I will be staying at *LOCATION.* Please suggest *ENTITY* cut off from *LOCATION.*
26	Any pleasant ideas of *ENTITY* far off the *LOCATION*? I might then be visiting *LOCATION.*
27	I came from *LOCATION* this afternoon. Any proposal for *ENTITY* not close to the *LOCATION*?
28	Does anyone have a suggestion for *ENTITY* distant from *LOCATION*? By the way, I came from *LOCATION* yesterday.
29	We will be staying near the *LOCATION.* Suggestions for *ENTITY* far from both *LOCATION* and *LOCATION* will be welcomed.
30	Any idea of *ENTITY* far away from *LOCATION* and *LOCATION*? I would then be visiting *LOCATION.*
31	Hi, I will be staying near the *LOCATION.* Could anyone propose *ENTITY* not very close to *LOCATION* and far from *LOCATION*?
32	Does anyone have suggestions for *ENTITY* far from *LOCATION* and also far from *LOCATION*? I will then be visiting *LOCATION* too.
33	Any good ideas of *ENTITY* far from *LOCATION* but close to *LOCATION* would be appreciated? Best Regards.
34	Anyone having ideas of *ENTITY* close to *LOCATION* but far from *LOCATION*?
35	Someone please advise *ENTITY* far from *LOCATION* but not very far from *LOCATION.*
36	Suggest *ENTITY* close to *LOCATION* but not in the neighborhood of *LOCATION.* Thank you so much!
37	Does anyone have good ideas of *ENTITY* far from *LOCATION* but near *LOCATION*? Regards.
38	Please suggest ideas of *ENTITY* in the neighborhood of *LOCATION* but far from *LOCATION.*
39	Could anyone advise *ENTITY* far from *LOCATION* but not too far from *LOCATION*?
40	Any nice ideas of *ENTITY* close to *LOCATION* but not in the neighborhood of *LOCATION.* Thanks!
41	Tomorrow, I would be coming to stay at *LOCATION.* Anyone having ideas of *ENTITY* close to *LOCATION* but far from *LOCATION*?
42	Please propose *ENTITY* far from *LOCATION* but not far from *LOCATION.* I will then be exploring *LOCATION.*
43	I came from *LOCATION* this evening. Any nice ideas for *ENTITY* far from *LOCATION* but close to *LOCATION* would be appreciated?
44	Suggest *ENTITY* close to *LOCATION* but not near *LOCATION.* Tomorrow, I will be leaving for *LOCATION.*
45	Yesterday, I came to stay at *LOCATION.* Any ideas of *ENTITY* close to *LOCATION* but far from *LOCATION*?
46	Suggestions of *ENTITY* far from *LOCATION* but not very far from *LOCATION.* I will then be moving to *LOCATION.*
47	I came from *LOCATION* today. Any good ideas for *ENTITY* far from *LOCATION* but near to *LOCATION* would be welcomed?
48	Advise *ENTITY* close to *LOCATION* but not close to *LOCATION.* I might be leaving for *LOCATION* soon.

A APPENDIX

This appendix is organized as follows.

- Section A.1 provides more details about the synthetic dataset and supplementary experimental information that includes additional tables referred to in the main paper on Spatial Reasoning.
- Section A.2 includes more results of the Location Tagger used in the end-task.
- Section A.3 contains supplementary experiments on Spatio-Textual Reasoning.
- Section A.4 gives details about the model hyper-parameters.

A.1 Synthetic Dataset

We create a simple synthetic dataset that is generated using linguistically diverse templates specifying spatial constraints and locations chosen at random from across 200,000 entities. These entities were sourced from the recently released Points-of-Interest (POI) recommendation task [9]. Each POI entity is labeled with its geo-coordinates apart from other meta-data such as its address, timings, etc. Further, each entity in a city has a specific type viz. Restaurant (R), Attraction (A) or Hotel (H). Table 10 shows the list of templates used for generating the dataset. These templates have been made to make the synthetic dataset reflective of real-world challenges. For instance, templates #41-#48 include the possibility of injecting *distractor locations*. To generate questions, $LOCATION and $ENTITY values are updated by randomly selecting values from the POI-set for each entity as described in the next section.

A.1.1 Dataset Generation. To generate a question, a city *c*, type *t* and a template *T* are chosen at random. The "ENTITY" token in each template is replaced by a randomly chosen *metonym* of the type *t*. Table 11 shows the list of metonyms for each type. Each instance of the "LOCATION" token is replaced by a randomly chosen entity from the city *c* and type *t*. The candidate set consists of the entities from the city *c* and type *t*. The entities used as location mentions

Table 11: List of metonyms for each entity type in the synthetic dataset

Entity type	Metonyms
R (Restaurant)	a restaurant, an eatery, an eating joint, a cafeteria, an outlet, a coffee shop, a fast food place, a lunch counter, a lunch room, a snack bar, a chop house, a steak house, a pizzeria, a coffee shop, a tea house, a bar room
H (Hotel)	a hotel, an inn, a motel, a guest house, a hostel, a boarding house, a lodge, an auberge, a caravansary, a public house, a tavern, an accomodation, a resort, a youth hostel, a bunk house, a dormitory, a flop house
A (Attraction)	an attraction, a tourist spot, a tourist attraction, a popular wonder, a sight-seeing place, a tourist location, a place of tourist interest, a crowd pleaser, a scenic spot, a popular landmark, a monument

are sampled without replacement and removed from the candidate set.

The gold answer entity is uniquely determined for each question based on its template. For example, consider a template T, "*I am staying at $A! Please suggest a hotel close to $B but far from $C.*" The score of a candidate entity X is given by $dist_T(X) = -(dist(X,B) - dist(X,C))$ (distances from B needs to be reduced, while distance from C needs to be higher). A is a distractor. The candidate with the $max(dist_T(X))$ in the universe is chosen as the gold entity for that question.

Each question further consists of 500 negative samples (35% hard, 65% soft). The negative samples are generated as a part of the gold generation process. A hard negative sample has a $dist_T(X)$ value closer to the gold as compared to a soft negative sample. We release the samples used for training along with the dataset for reproducibility.

A.1.2 Template classes. We create templates (Table 10) that can be broadly divided into three different categories based on whether the correct answer entity is expected to be: (1) close to one or more locations [1-16] (2) far from one or more locations [17-32] (3) close to some and far from others (combination) [33-48]. To make the task more reflective of real-world challenges we also randomly insert a *distractor* location that does not need to be reasoned. The second-half for each category (i.e. [9-16], [25-32], and [41-48]) consists of templates that have a distractor locative reference. Further, for the close (or far) category, the templates could contain one location ([1-4] + [9-12]) or two locations ([5-8] + [13-16]) that need to be reasoned for close (or far).

A.1.3 Results. We use the following models in our experiments:(i) SPNet (ii) SPNet without (w/o) DRL (iii) BERT-SPNet (iv) BERT-SPNet without (w/o) DRL. Models without DRL use the final hidden states of the Question Encoder and a series of down-projecting feed-forward layers to generate the final score.

We study our models' performance using $Hits@N$ (N=3,5,30) which requires that any one of the top-N answers be correct, Mean Reciprocal Rank (MRR), and the average distance of the top-3 ranked answers from the gold-entity $Dist_g$. Table 12 summarizes the results on the test set.

A.1.4 Error Analysis. Tables 13 and 14 show the effect of candidate search space and the number of location mentions in the question on the performance of the SPNet Model.

Table 12: Results of the Spatial-reasoning network on the synthetic data test set

Models	Hits@3	Hits@5	Hits@30	MRR	Dg
	Close to Set X				
SPNet w/o DRL	62.60	66.00	79.00	0.608	2.88
SPNet	90.20	**92.80**	**97.60**	0.873	0.86
BERT SPNet w/o DRL	63.60	67.60	82.60	0.616	3.68
BERT SPNet	**91.40**	**92.80**	97.20	**0.896**	**0.78**
	Far from Set X				
SPNet w/o DRL	89.00	90.80	96.40	0.858	15.24
SPNet	**98.00**	**98.40**	**99.20**	0.975	13.88
BERT SPNet w/o DRL	90.80	92.00	95.80	0.881	15.32
BERT SPNet	97.80	98.00	98.80	**0.978**	**13.87**
	Combination				
SPNet w/o DRL	23.40	28.00	50.60	0.229	9.72
SPNet	52.80	60.20	82.00	0.486	3.90
BERT SPNet w/o DRL	26.80	32.60	59.00	0.242	12.96
BERT SPNet	**59.20**	**65.80**	**86.20**	**0.551**	**3.02**
	Aggregate				
SPNet w/o DRL	58.33	61.60	75.33	0.565	9.28
SPNet	80.33	83.80	92.93	0.778	6.21
BERT SPNet w/o DRL	60.40	64.07	79.13	0.579	10.65
BERT SPNet	**82.80**	**85.53**	**94.07**	**0.808**	**5.89**

Table 13: Performance of SPNet decreases with increase in universe size.

	Correctly Answered		Incorrectly Answered	
Search Space size	**Questions**	**Percentage**	**Questions**	**Percentage**
0-200	318	26.39%	42	14.24%
200-500	417	34.61%	83	28.13%
500-1000	221	18.34%	53	17.97%
1000-5000	178	14.77%	82	27.80%
5000-20000	71	5.89%	35	11.86%

Table 14: Performance of SPNet decreases with increase in the number of location mentions in the question.

	Correctly Answered		Incorrectly Answered	
# Location-Mentions	**Questions**	**Percentage**	**Questions**	**Percentage**
1	233	19.34%	0	0.00%
2	671	55.69%	126	42.71%
3	301	24.98%	169	57.29%

A.2 Location Tagger

In order to get mentions of locations in questions, we manually label a set of 425 questions from the training set for location mentions. We then use a BERT-BiLSTM CRF [8] based tagger trained on this set to label locations. Table 15 describes the performance of the tagger on an unseen set of 75 questions.

Table 15: Performance of the BERT-BiLSTM CRF for tagging locations on a small set of 75 unseen questions.

	Precision	Recall	F1
Micro Average	87.59	87.56	87.58
Macro Average	88.24	87.83	88.03

A.3 Spatio-textual Reasoning Network

The Spatio-Textual Reasoning Network consists of three components (i) Geo-Spatial Reasoner (ii) Textual Reasoner (iii) Joint Scoring Layer.

Training: We train the joint model using max-margin loss teaching the network to score the correct-answer higher than a negatively sampled candidate entity. Model parameters are described in the next section.

A.3.1 Results. Similar to Contractor et al. [9] we also experiment on this dataset by employing a neural method to reduce the search space [25] before using the CrQA textual-reasoner to re-rank only the top-30 selected candidates (pipeline referred to as CsrQA). Unlike CrQA, which uses two levels of attention between question and review sentences to score candidate entities CsQA does not reason deeply over the text. It compares elements of a question with different parts of a review document to aggregate relevance for scoring. Local and distributed representations are used to capture lexical and semantic features.

We report some experiments (Table 16) using this model referred to as CsQA and compare it with CsrQA and spatio-textual CsrQA. As can be seen re-ranking with SD or SPNet does not help the system. An interesting direction of future work could thus, be to augment general-purpose neural-IR methods such as Duet used by CsQA with spatial-reasoning. Another interesting approach could be to extend ideas from existing Graph-neural network based approaches, such as NumNet [27]. Each entity could be viewed as a node in a graph for reasoning but we note that methods will need to be made more scalable for them to be useful. The entity space (and thus nodes in the graph) would run into thousands of nodes per question making current message-passing based inference methods prohibitively expensive.

Table 16: Comparison of re-ranking models operating on a reduced search space returned by CsQA on Location Questions (ii) Comparison with current state-of-the-art CsrQA on the full task.

	Location Questions				
Models	**Hits@3**	**Hits@5**	**Hits@30**	**MRR**	**Dg**
CsQA	15.84	20.26	**51.47**	0.149	2.61
CsQA → SD	11.34	17.26	**51.47**	0.118	**2.18**
CsQA → LocNet	8.38	13.72	**51.47**	0.097	2.27
CsrQA	19.89	26.43	**51.47**	0.168	2.70
Spatio-textual CSRQA	**21.36**	**28.36**	**51.47**	**0.183**	2.27
	All Questions				
CsrQA	21.45	28.21	**52.65**	0.186	2.47
Spatio-textual CSRQA	**22.41**	**28.99**	52.65	**0.193**	**2.32**

A.4 Model settings

A.4.1 Experiments on Synthetic Dataset. The hyperparameters for the best performing configurations of all models were identified through manual testing on the validation set (Table 17). The models were trained on a 2x NVIDIA K40 (12GB, 2880 CUDA cores) GPU on a shared cluster.

Table 17: Hyperparameter settings for experiments on the synthetic dataset

Hyperparameter	**Value**
Negative samples	40
Batch size	20
Optimizer	Adam
Loss	MarginRankingLoss
Margin	0.5
Max no. of epochs	15
GRU Input dimension	131
GRU Output dimension	32
DRL Block Layer 1	64 (Input) 64 (Output)
DRL Block Layer 2	64 (Input) 64 (Output)
DRL Block Layer 3	64 (Input) 64 (Output)
DRL Block Layer 4	64 (Input) 1 (Output)

The BERT models were trained with a learning rate of 0.0002 whereas the non-BERT models with a learning rate of 0.001.

A.4.2 Spatio-textual Reasoning Network. The hyper-parameters for the best performing configuration were identified through manual testing on the validation set (Table 18). The Spatio-Textual Reasoner was trained on 4 K-80 GPUs on a shared cluster.

Table 18: Hyperparameters used for experiments on the end-task

Hyperparameter	**Value**
Word embeddings size	128
Dropout	0.2
Optimizer	Adam
Loss	Hinge Loss
Margin	1.0
Batch Size	200
SPNet GRU input dimension	131
SPNet GRU output dimension	256
Textual GRU input dimension	128
Textual GRU output dimension	256
DRL Block Layer 1	512 (Input) 256 (Output)
DRL Block Layer 2	256 (Input) 256 (Output)
DRL Block Layer 3	256 (Input) 128 (Output)
DRL Block Layer 4	128 (Input) 128 (Output)
DRL Block Layer 5	128 (Input) 50 (Output)
DRL Block Layer 6	50 (Input) 10 (Output)
DRL Block Layer 7	10 (Input) 1 (Output)
α,β FF Linear Layer 1	256 (Input) 50 (Output)
α,β FF Linear Layer 2	50 (Input) 50 (Output)
α,β FF Linear Layer 3	50 (Input) 10 (Output)
α,β FF Linear Layer 4	10 (Input) 10 (Output)
α,β FF Linear Layer 5	10 (Input) 10 (Output)
α,β FF Linear Layer 6	10 (Input) 2 (Output)