

**QA FOR TOURISM: ANSWERING  
RECOMMENDATION AND  
COMPARISON QUESTIONS**

**DANISH CONTRACTOR**



**AMARNATH AND SHASHI KHOSLA SCHOOL OF IT  
INDIAN INSTITUTE OF TECHNOLOGY DELHI  
NOVEMBER 2021**

©Indian Institute of Technology Delhi - 2021  
All rights reserved.

# QA FOR TOURISM: ANSWERING RECOMMENDATION AND COMPARISON QUESTIONS

by

DANISH CONTRACTOR

AMARNATH AND SHASHI KHOSLA SCHOOL OF IT

Submitted

in fulfillment of the requirements of the degree of Doctor of Philosophy

to the



INDIAN INSTITUTE OF TECHNOLOGY DELHI  
NOVEMBER 2021

DEDICATED TO  
*My parents - Yasmin and Osman*

# Certificate

This is to certify that the thesis titled **QA For Tourism: Answering Recommendation and Comparison Questions** being submitted by **Danish Contractor** for the award of **Doctor of Philosophy** in Department of Computer Science and Engineering is a record of bona fide work carried out by him under my guidance and supervision at the Amarnath and Shashi Khosla School of IT, Indian Institute of Technology Delhi. The work presented in this thesis has not been submitted elsewhere, either in part or full, for the award of any other degree or diploma unless otherwise stated explicitly. In particular, work done in Chapters 3, 4 and 5 were done jointly with undergraduate students. In each case, the part done by the collaborators appeared in their respective Bachelor's theses.

**Mausam**

Professor

Department of Computer Science and Engg.

Indian Institute of Technology Delhi

New Delhi- 110016

**Parag Singla**

Associate Professor

Department of Computer Science and Engg.

Indian Institute of Technology Delhi

New Delhi- 110016

# Acknowledgments

I would not have been able to pursue my PhD, were it not for the constant nurturing support of my family - especially my mother Yasmin, who has been so patient and understanding throughout my doctoral journey. A special thanks also goes out to my sister, Seher, for always being there for me. It was only due to the supportive and enabling environment at home that allowed me to dedicate most of my waking hours to my work.

I would like to extend my deepest gratitude to my advisors - Mausam and Parag Singla. Thank you for having faith in me and giving me the freedom to work on problems that I found interesting. I am grateful for the encouragement, the continuous feedback, guidance and support throughout. You've both been a source of inspiration to me.

During the entire duration of my PhD, I have been a full-time employee at IBM Research, which continues to be a great place to work. I have learnt a lot from my colleagues and my PhD would not have been possible without the support of my friends and co-workers at IBM. I would especially like to thank my mentor, Venkat Subramaniam who gave me the first opportunity to learn and grow as a researcher and then also motivated me to pursue a PhD. I would like to thank my managers over the years - Sachin Joshi, Bikram Sengupta, Renuka Sindhgatta, Karthik Sankarnaryanan for trusting me and giving me the time, resources and flexibility to balance my work commitments during the course of my PhD. Thanks goes to my colleagues and teammates Dinesh Raghu, Vineet Kumar, Gaurav Pandey, Dhiraj Madan, Sumit Negi, Sumit Bhatia, Sachin Joshi for their support and helpful technical feedback.

I would like to thank and acknowledge my fantastic student collaborators Poojan Mehta, Barun Patra, Krunal Shah, Aditi Partap and Shashank Goel, all of whom have played a crucial role in shaping this work. Thank you Happy Mittal, Ankit Anand, Dinesh Khandelwal for always being a phone call away and helping me progress through the milestones of the PhD.

Thank you Azalenah, for all the 'thesis snacks' and for painstakingly helping me proof-read the entire (boring) thesis! Thank you Saumya Saxena, Aparajita Bharti, Naaz Mustafa and Wajida Contractor for helping me with translations of the Abstract. A shout-out to my amazing circle of friends - *'Thread'*, *'Doston'*, *'X G'*, *'Cantab'*, *'Kabootars of Trafalgar'*, *'Dim Bits'*, *'Bombay Kids'*, *'Rohit Kumar & Kids'* - thank you for all the joy, jokes and laughter!

# Abstract

Travellers often post questions online to seek personalized travel recommendations by describing their preferences and constraints with respect to locations, points of interests, budget, etc. They also, at times, post queries asking for comparisons between cities, tourism sites, etc when making their travel plans. In this thesis we study the novel tasks of answering such *recommendation* and *comparison* questions from the tourism domain.

We focus our attention on a class of recommendation questions that seek entities. We refer to them as Multi-sentence entity-seeking recommendation questions (MSRQs) i.e., questions that expect one or more entities as an answer. In the tourism domain, such entities can occur in the form of Points-of-Interest (POIs); e.g, names of hotels, restaurants, tourist sites. We answer entity-seeking recommendation questions in two settings: (i) QA with intermediate annotations (ii) QA without intermediate annotations. In each setting we formulate a new problem and create new datasets which we hope will help further research in QA. In the first setting, we develop a pipelined model which breaks down the task of question-answering into a question-parsing task followed by knowledge-base querying. Learning a question-parser requires large amounts of training data and we overcome this challenge by employing a constraint driven learning framework that uses a small set of expert-annotated questions, along with a larger set of crowd-sourced partially-annotated questions.

In contrast to the first setting, in the second approach we use a collection of reviews to directly answer questions, without explicitly parsing questions. Answering such questions poses novel challenges of reasoning at scale, since review collections for each entity can be very large, noisy, contain subjective opinions, and each question can have thousands of entities to choose from to return as ‘possible answers’. In response, we present a cluster-retrieve-rerank architecture that helps address some of these challenges. It first clusters review text for each entity to identify exemplar sentences describing an entity. It then uses a scalable neural information retrieval (IR) module to select a set of potential entities from the large candidate set. A reranker uses a deeper attention-based architecture to pick the best answers from the selected entities. Additionally, in order to accommodate reasoning over physical locations of entities, we extend this work by developing a joint spatio-textual model. We 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. We combine the spatial-reasoner with the textual QA system to develop a joint spatio-textual QA model. We demonstrate that our joint spatio-textual model performs significantly better than models employing only spatial or textual reasoning.

Lastly, we also study the problem of answering comparison questions. We define a novel task of generating entity comparisons from textual corpora in which each document describes one entity at a time. We generate entity comparisons in a tabular form in which attribute-value phrases, opinion phrases, and other descriptions are clustered and organized topically, thus, allowing for direct comparisons. Our tabular summaries balance information about the entities being compared and in our user studies we find that users strongly preferred balanced clusters, and acquire as much information about the entities, by using the tables, as they do using articles.

## सार

यात्री अक्सर अपनी पसंद, रुचियों के स्थल , बजट आदि के ज़रूरत के संबंध में व्यक्तिगत यात्रा सुझाव प्राप्त करने के लिए ऑनलाइन प्रश्न पोस्ट करते हैं। वे कभी-कभी अपने प्रश्न बनाते समय शहरों, पर्यटन स्थलों आदि के बीच तुलना करने के लिए भी पूछते हैं। इस थीसिस में हम पर्यटन से इस तरह के सुझाव और तुलनात्मक प्रश्नों के उत्तर देने के नए कार्यों का अध्ययन करते हैं। हम अपना ध्यान सुझाव प्रश्नों के एक वर्ग पर केंद्रित करते हैं जो अहम स्थलों की तलाश करते हैं। हम उन्हें बहु-वाक्य एंटीटी-खोज सुझाव प्रश्न के रूप में संदर्भित करते हैं, अर्थात्, ऐसे प्रश्न जो उत्तर के रूप में एक या अधिक स्थलों की अपेक्षा करते हैं। पर्यटन क्षेत्र में, ऐसी स्थल 'रुचि के स्थल' (पीओआई) के रूप में हो सकती हैं; जैसे, होटल, रेस्तरां, पर्यटन स्थलों के नाम। हम सुझाव चाहने वाले प्रश्नों का उत्तर दो सेटिंग्स में देते हैं: (i) मध्यवर्ती एनोटेशन के साथ प्रश्नोत्तर (ii) मध्यवर्ती एनोटेशन के बिना प्रश्नोत्तर। प्रत्येक सेटिंग में हम एक नई दिशा में अन्वेषण करते हैं और नए डेटासेट बनाते हैं जो हमें उम्मीद है कि प्रश्नोत्तर में आगे के शोध में मदद करेगा। अंत में, हम तुलनात्मक प्रश्नों के उत्तर देने का भी प्रयत्न करते हैं। हम टेक्स्ट कॉर्पोरा से वस्तु तुलना उत्पन्न करने का एक नया कार्य परिभाषित करते हैं जिसमें प्रत्येक दस्तावेज़ एक समय में एक एंटीटी का वर्णन करता है। हम एक सारणीबद्ध रूप में एंटीटी तुलना उत्पन्न करते हैं जिसमें विशेषता-मूल्य वाक्यांश और राय वाक्यांश होते हैं। हमारे सारणीबद्ध सारांश तुलना जानकारी को संतुलित करते हैं और हमारे उपयोगकर्ता अध्ययनों में हम पाते हैं कि वह संतुलित समूहों को दृढ़ता से पसंद करते हैं।

## تلخیص

مسافرین اکثر اپنی ترجیحات اور تحدیدات بیان کرتے ہوئے مخصوص مقامات، پسندیدہ مقامات، بجٹ وغیرہ سے متعلق سفری تجاویز حاصل کرنے کیلئے آن لائن سوالات پوسٹ کرتے ہیں۔ اور ساتھ ہی سفر کی منصوبہ بندی کے وقت وہ شہروں، سیاحتی مقامات کے درمیان موازنہ سے متعلق سوالات بھی پوسٹ کرتے ہیں۔ ان مقالہ جات میں ہم شعبہ سیاحت سے متعلق اس قسم کی تجاویز اور موازنہ کے جوابات دینے کے نئے کام کا جائزہ لینگے۔ ہم اپنی توجہ تجاویز اور سوالات پر مرکوز کئے ہوئے ہیجوا کسی چیز کی طالب ہیں۔ ہم اسے ملٹی سینٹنس انٹیٹی سیکنگ تجاویز سوالات (MSRQs) قرار دیتے ہیں مثلاً ایسے سوالات جو ایک یا ایک سے زائد جوابات چاہتے ہیں۔ اس قسم کے انٹی ٹیز شعبہ سیاحت میں پوائنٹس آف انٹرسٹ کی (POIs) طرح ہوتے ہیں، مثلاً ہوٹلس، رسٹورنٹ، سیاحتی مقامات۔ ہم انٹیٹی سیکنگ تجاویز اور سوالات کے دو ترتیب میں جواب دیتے ہیں انٹرمیڈیٹ QA (i): تشریحات کیساتھ (ii) QA انٹرمیڈیٹ تشریحات کے بغیر۔ ہر ترتیب میں ہم ایک نیا مسئلہ وضع کرتے ہیں اور نئے ڈیٹا سیٹس بناتے ہیں جس سے امید ہے کہ ہمیائندہ QA کی تحقیق میمزید مدد ملے گی۔ آخر میں ہم موازنہ کے سوالات کا جواب دینے سے متعلق مسئلہ کی بھی اسٹڈی کرتے ہیں۔ ہم ٹیکسٹل کارپورا سے انٹیٹی موازنہ کو نکالنے کے کام کو وضع کرتے ہیں جس میں ہر ڈاکیومنٹ ایک وقت میں ایک انٹیٹی کو بیان کرتا ہے۔ ہم انٹیٹی موازنہ کو جدولی شکل میں تیار کرتے ہیں جس میں اٹریبوٹ والیوفریسس، تجاویز کے فریسس اور دیگر وضاحتیں کلسٹرڈ اور لفظی اعتبار سے منظم ہوتی ہے۔ جو براہ راست موازنہ کی اجازت دیتی ہے۔ انٹیٹس کے بارے میں ہمارے جدولی خلاصے، متوازن انفارمیشن جسکا موازنہ کیا جاتا ہے۔ ہمارے استفادہ کنندوں کی تحقیق میں ہم نے یہ بھی پایا کہ استفادہ کنندے متوازن کلسٹرس کو زیادہ ترجیح دیتے ہیں۔

# Contents

<b>Certificate</b>	<b>iii</b>
<b>Acknowledgments</b>	<b>v</b>
<b>Abstract</b>	<b>vii</b>
<b>I Prologue</b>	<b>1</b>
<b>1 Introduction</b>	<b>3</b>
1.1 Types of Questions in a Travel forum . . . . .	5
1.2 Scope of Research . . . . .	6
1.3 Contributions . . . . .	7
1.3.1 Answering Recommendation Questions . . . . .	7
1.3.2 Answering Comparison Questions . . . . .	10
1.4 Thesis Outline . . . . .	11
<b>2 Background &amp; Related Work</b>	<b>13</b>
2.1 Sequence Modeling and Tagging . . . . .	13
2.1.1 Recurrent Neural Networks . . . . .	14
2.1.2 Transformer Networks . . . . .	17
2.1.3 Conditional Random Fields for Sequence Tagging . . . . .	18
2.2 Topic Modeling . . . . .	20
2.2.1 LSI: Latent Semantic Indexing . . . . .	20
2.2.2 pLSI: Probabilistic Latent Semantic Indexing . . . . .	21
2.2.3 Latent Dirichlet Allocation . . . . .	22
2.2.4 Gaussian Mixture Models and Clustering . . . . .	23
2.3 Question Answering Tasks . . . . .	24
2.3.1 QA using Structured & Semi-structured Knowledge . . . . .	24
2.3.2 QA using Unstructured Knowledge . . . . .	25

2.3.3	Attention in Neural Question Answering . . . . .	27
<b>II</b>	<b>Recommendation Questions</b>	<b>29</b>
<b>3</b>	<b>QA with Intermediate Annotations</b>	<b>31</b>
3.1	Contributions . . . . .	33
3.2	Overview . . . . .	33
3.3	Related Work . . . . .	34
3.3.1	Question Answering Systems . . . . .	35
3.3.2	Question Parsing . . . . .	37
3.3.3	Neural Semantic Parsing . . . . .	37
3.3.4	Summary . . . . .	38
3.4	Semantic Labels for MSRQs . . . . .	38
3.5	MSRQ Semantic Labeling . . . . .	39
3.5.1	Features . . . . .	40
3.5.2	Constraints . . . . .	42
3.5.3	Partially labeled data . . . . .	45
3.5.4	Crowd-sourcing Task . . . . .	45
3.5.5	Training with partially labeled posts . . . . .	47
3.6	Evaluation . . . . .	48
3.6.1	Dataset . . . . .	48
3.6.2	Methodology . . . . .	48
3.6.3	Results . . . . .	49
3.7	Answering System . . . . .	51
3.8	Understanding MSRQs in another domain . . . . .	56
3.9	Summary . . . . .	57
<b>4</b>	<b>QA without Intermediate Annotations</b>	<b>59</b>
4.1	Contributions . . . . .	61
4.2	Data Collection . . . . .	61
4.2.1	Answer Extraction . . . . .	62
4.2.2	Filtering of Silver Answer Entities . . . . .	63
4.2.3	Qualitative Study: Data . . . . .	64
4.2.4	Data Characteristics . . . . .	66
4.3	Problem Statement . . . . .	67
4.4	Related Work: QA & IR . . . . .	67
4.4.1	Duet – a Neural IR Network . . . . .	69

4.5	The Cluster-Select-Rerank Model . . . . .	70
4.5.1	<i>Cluster: Representative Entity Document Creation</i> . . . . .	71
4.5.2	<i>Select: Shortlisting Candidate Answers</i> . . . . .	71
4.5.3	<i>Rerank: Answering over Selected Candidates</i> . . . . .	72
4.6	Evaluation . . . . .	74
4.6.1	Models for comparison . . . . .	74
4.6.2	Metrics for Model evaluation . . . . .	76
4.6.3	Results . . . . .	77
4.6.4	Sampling Strategies for Curriculum Learning . . . . .	80
4.7	Summary . . . . .	81
<b>5</b>	<b>Improving QA with Spatio-Textual Reasoning</b>	<b>83</b>
5.1	Contributions . . . . .	84
5.2	Related Work . . . . .	85
5.3	Spatio-Textual Reasoning Network . . . . .	86
5.3.1	Geo-Spatial Reasoner . . . . .	87
5.3.2	Textual-Reasoning Sub-network . . . . .	90
5.3.3	Joint Scoring Layer . . . . .	90
5.4	Evaluation . . . . .	91
5.4.1	Detailed Study: Geo-Spatial Reasoner . . . . .	91
5.4.2	Spatio-Textual Reasoning Network . . . . .	96
5.5	Summary . . . . .	102
<b>III</b>	<b>Comparison Questions</b>	<b>103</b>
<b>6</b>	<b>Automated Entity Comparison</b>	<b>105</b>
6.1	Contributions . . . . .	107
6.2	Related Work . . . . .	107
6.3	Task & System Description . . . . .	108
6.4	Architecture . . . . .	109
6.4.1	Information Extraction . . . . .	109
6.4.2	Building Clusters for Comparison . . . . .	112
6.5	Evaluation . . . . .	116
6.5.1	Evaluation of Clustering Algorithms . . . . .	117
6.5.2	Value of Comparison Tables . . . . .	118
6.6	Summary . . . . .	120

<b>IV</b>	<b>Epilogue</b>	<b>121</b>
<b>7</b>	<b>Conclusion &amp; Future Work</b>	<b>123</b>
7.1	Improving Joint-Reasoning . . . . .	124
7.1.1	Question Answering . . . . .	124
7.1.2	Clustering . . . . .	125
7.2	Improving Textual Reasoning . . . . .	125
7.3	Task Extensions . . . . .	127
7.4	Extension to other domains . . . . .	128
<b>A</b>	<b>POI-Recommendation Dataset Statistics</b>	<b>131</b>
<b>B</b>	<b>Joint Spatio-Textual Reasoning</b>	<b>137</b>
B.1	Artificial Dataset . . . . .	137
B.1.1	Dataset Generation . . . . .	137
B.1.2	Template classes . . . . .	139
B.2	Model settings . . . . .	139
B.2.1	Experiments on artificial dataset . . . . .	139
B.2.2	Spatio-textual Reasoning Network . . . . .	140
<b>C</b>	<b>Pairs used for Comparisons</b>	<b>141</b>
<b>D</b>	<b>Answering Comparison Questions: Screenshots of Crowd-worker Tasks</b>	<b>145</b>
<b>E</b>	<b>System Outputs</b>	<b>151</b>
E.1	Recommendation Questions . . . . .	151
E.1.1	Example 1: Restaurant recommendation with location constraints (Correct answer returned) . . . . .	151
E.1.2	Example 2: Restaurant recommendation with location constraints (Correct answer returned) . . . . .	156
E.1.3	Example 3: Hotel recommendation with location constraints and budgetary constraints (Partially correct answer returned) . . . . .	160
E.1.4	Example 4: Hotel Recommendation (Correct answer returned) . . . . .	163
E.1.5	Example 5: Restaurant Recommendation (Incorrect answer returned) . . . . .	165
E.2	Comparison Questions . . . . .	166
	<b>Bibliography</b>	<b>173</b>
	<b>List of Publications</b>	<b>201</b>

Biography

203

Variable	Definition
$i, j, k, l$	Locally declared indexing variables
$x, y, z$	Locally declared variables to define functions
$\mathbb{M}$	Training Set
$\mathbb{C}$	Set of cities
$\mathbb{E}$	Set of entities
$\mathbb{U}$	Set of partially labeled posts
$\mathbb{P}$	Set of POI types supported; $\mathbb{P} = \{\text{hotel, attraction, restaurant}\}$
$p$	POI type $p \in \mathbb{P}$
$q$	a user question
$q$	Embedding representation of $q$
$e$	An entity $e \in \mathbb{E}$
$e$	Embedding representation of $e$
$\mathbf{E}_e$	Matrix consisting of sentence embeddings for an entity $e$
$\hat{e}_q$	Question-aware embedding representation of $e$
$\phi$	CRF Feature
$\omega$	CRF Feature Weight
$\rho_k$	Weight associated with $k^{\text{th}}$ constraint in Constraint Conditional Modeling (CCM)
$C_k$	Violation score associated with the $k^{\text{th}}$ constraint in CCM
$\gamma$	Weight to control importance given to partially labeled posts in CCM
$\mathbf{A}$	Attention weights (matrix) for an encoded sequence
$\mathbf{H}$	Matrix consisting of hidden states of an encoded sequence
$\mathbf{W}_E$	Weight matrix for computing question-entity attention
$\mathbf{A}_E$	Attention weights (matrix) for generating entity embeddings
$\mathbf{w}$	Distance weight vector
$w_i^d$	Distance weight of the $i^{\text{th}}$ location-mention
$d_k$	Distance of an entity from the $k^{\text{th}}$ location mention in a question
$\mathbf{d}'$	Distance vector
$\mathbf{A}_l, b_l$	Weight matrix and bias term respectively for any feed-forward block at layer $l$
$\mathbf{B}$	Vector of position indices in question with $B$ label after $B - I$ encoding
$\mathcal{S}_T$	Textual Reasoning Score
$\psi_T$	Scaling weights for $\mathcal{S}_T$
$\mathcal{S}_L$	Spatial Reasoning Score
$\psi_L$	Scaling weights for $\mathcal{S}_L$
$\mathcal{S}$	Entity Relevance Score
$\sigma$	Sigmoid function
$\bar{\rho}$	Spearman's rank coefficient
$\alpha, \beta$	Weights for joint scoring in Spatio-Textual Reasoning
$\Theta$	Parameters of a model
$\eta$	Regularization Weight for EB G-pLSA

Table 1: List of variables

# List of Figures

1.1	Travel aggregator website Kayak.com allows searching for multi-destination flight options, hotels, deals, etc. . . . .	4
1.2	A question posted on a popular travel forum website - TripAdvisor.com along with responses from forum users. . . . .	5
2.1	Rolled computational graph depicting an RNN. Figure adapted from [Goodfellow et al., 2016]. . . . .	14
2.2	Conditional Random Field depicted as a graphical model . . . . .	18
2.3	Latent Semantic Indexing . . . . .	20
2.4	Probabilistic Latent Semantic Indexing – the figure uses <i>Plate Notation</i> to depict the graphical model. Here, ‘circles’ correspond to random variables and the ‘rectangles’ (plates) correspond to repetitions of random variables. Shaded circles denote observed variables while un-shaded circles denote unobserved (latent) variables. . . . .	20
2.5	Latent Dirichlet Allocation . . . . .	22
3.1	An entity-seeking MSRQ and annotated with our semantic labels . . . . .	31
3.2	Schematic Representation of the QA system . . . . .	34
3.3	BERT BiLSTM CCM with features for sequence labeling. . . . .	42
3.4	Snippet of the second questionnaire given to AMT workers . . . . .	46
4.1	Entity Answers are extracted from forum post responses to generate QA Pairs. Entities marked in red indicate false positive extractions. Each entity in our collection has an ID of the form <city_id >_<POI type>_<number>. The dataset has three classes of POIs - restaurants (R), attractions (A) and hotels (H). Example forum question from <a href="https://bit.ly/2zIxQpj">https://bit.ly/2zIxQpj</a> adapted for illustration. . . . .	62
4.2	Human Intelligence Task (HIT) set up on Amazon Mechanical Turk to clean test and validation sets. . . . .	64

4.3	The Duet retrieval model [Mitra et al., 2017, Mitra and Craswell, 2019] .	69
4.4	Representative documents created from Bag-of-Reviews entity documents, using clustering. . . . .	71
4.5	Reasoning network used to re-rank candidates shortlisted by the Duet model. . . . .	73
4.6	Entity class-wise break-up of the number of times (and %) a correct answer was within the top-3 ranks binned based on the size of candidate search space (<100, 100-1000, 1000+ entities) (X-axis). . . . .	78
5.1	A sample POI recommendation question from our dataset created in Chapter 4. The answers correspond to POI IDs of the form <city_id >_<POI type>_<number>. . . . .	83
5.2	Spatio-Textual reasoning network consisting of (i) Geo-Spatial Reasoner (ii) Textual-Reasoning subnetwork (iii) Joint Scoring Layer . . . . .	88
5.3	Sample questions from the artificial dataset. The dataset has questions from three categories: (1) close to set X, (2) far from set X (3) Combination. . . .	91
5.4	Probing study of the Distance Reasoning Layer (DRL) using the question: “ <i>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’?</i> ”. The coloured boxes indicate the relative magnitude of weights assigned; each candidate entity assigns a <i>higher</i> 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 . . .	94
5.5	Performance of SPNET decreases with increase in universe size. . . . .	95
5.6	Performance of SPNET decreases with increase in the number of location mentions in the question. . . . .	96
6.1	Sample comparison for two cities - Granada (Spain) and New York City (United States) generated using our system. A quick look reveals that that both cities have a nice set of museums and gardens to visit, while palaces and courtyards are only in Granada. Granada’s art and architecture are more ornamental, whereas New York’s might be more contemporary. . .	106
6.2	Information Extraction pipeline based on a seed list generated using LDA	110
6.3	Three alternative clusterings (a), (b), (c) for descriptive phrases from two cities – each color is a different city. We prefer clusters shown in (c) as they balance information from both entities . . . . .	112

---

6.4	Plate Notation of (i) Standard Gaussian Mixture Model (ii) Gaussian pLSA (and entity balanced Gaussian pLSA) . . . . .	113
6.5	Sample comparison for two movies - Batman (1989) and Gandhi (1982), generated using our system. . . . .	118
7.1	Example of a multi-modal recommendation question. Answering this question, requires understanding the information encoded in the images. Question and image source: <a href="https://www.houzz.com/discussions/5643190/best-garage-floor-epoxy#n=15">https://www.houzz.com/discussions/5643190/best-garage-floor-epoxy#n=15</a> . . . . .	129
D.1	Sample task screenshot where users were shown the comparison tables before writing summaries. A live timer displayed current time left for task. Screenshot truncated for ease of presentation. . . . .	146
D.2	Sample task screenshot where users were asked to write summaries after viewing the comparison table. A live timer displayed current time left for task. Screenshot truncated for ease of presentation. . . . .	147
D.3	Sample task screenshot where users were shown the full articles before writing summaries. A live timer displayed current time left for task. Screenshot truncated for ease of presentation. . . . .	148
D.4	Sample task screenshot where users were asked to write summaries after viewing the full articles. A live timer displayed current time left for task. Screenshot truncated for ease of presentation. . . . .	149
D.5	Sample task screenshot where users were asked compare written summaries. Screenshot truncated for ease of presentation. . . . .	150
E.1	Comparison generated between two cities - Singapore and Philadelphia. Truncated for ease of presentation. . . . .	167
E.2	Comparison generated between two cities - Singapore and Abu Dhabi. Truncated for ease of presentation. Notice the different topical organization; in contrast to the previous example, there is a finer cluster around beaches. . . . .	168
E.3	Comparison generated between two cities - Singapore and Kuala Lumpur. Truncated for ease of presentation. Notice the different topical organization; in contrast to the previous example, there is a finer cluster around colonial buildings and there is a comparative cluster for beaches and parks. . . . .	169
E.4	Comparison generated between Rome and Goa. Truncated for ease of presentation. Amongst others, notice the cluster related to beaches and parks. . . . .	170

E.5 Comparison generated between two cities - Rome and Jerusalem. Truncated for ease of presentation. In contrast to the previous comparison, notice clusters related to Islamic and Jewish art emerge. . . . . 171

E.6 Comparison generated between two cities - Rome and Hyderabad. Truncated for ease of presentation. Notice the first cluster related to water bodies and the cluster related to temples and roman architecture that emerges in the comparison. . . . . 172

# List of Tables

1	List of variables . . . . .	xvi
3.1	Related work: Question Answering . . . . .	35
3.2	Regular Expressions of POS-based patterns used to create indicator features for <i>entity.type</i> tokens. We ignore <i>WP</i> tags when the tag is associated with ‘ <i>who</i> ’. . . . .	41
3.3	Agreement for <i>entity</i> labels on AMT . . . . .	47
3.4	Sequence tagger <i>F1</i> scores using CRF with all features (feat), CCM with all features & constraints, and partially-supervised CCM over partially labeled crowd data. The second set of results mirror these settings using a bi-directional LSTM CRF. Results are statistically significant (paired t-test, p value<0.02 for aggregate <i>F1</i> for each CRF and corresponding CCM model pair). Models with “PS” as a prefix use partial supervision. . . . .	50
3.5	Feature ablation study using a vanilla CRF model. . . . .	50
3.6	(i) Precision and Recall of <i>entity.type</i> with and without CCM inference. . . . .	51
3.7	Performance of negation detection using gold sequence labels, and system generated labels . . . . .	53
3.8	QA task results using the Google Places web API as knowledge source. . . . .	53
3.9	Some sample questions from our test set and the answers returned by our system. Answers in <b>green</b> are identified as correct while those in <b>red</b> are incorrect. . . . .	54
3.10	Classification of errors made by our MSRQ-labels based answering system (using Google Places web API as knowledge source) . . . . .	55
3.11	Labeling performance for Book recommendation questions (paired t-test, p value<0.01 for aggregate <i>F1</i> in vanilla CRF and CCM model pairs & BiLSTM CRF and CCM model pairs). . . . .	57
4.1	QA Pairs in train, validation and test sets . . . . .	65
4.2	Knowledge source consisting of 216,033 entities and their reviews . . . . .	66

4.3	Classification of Questions. (%) does not sum to 100, because questions may exhibit more than one feature. . . . .	66
4.4	Related datasets on Machine reading/QA and their characteristics. Unlike other existing datasets, our task requires us to reason over <i>opinions</i> . For reading comprehension tasks, the document containing the actual answer may not always be known. *“docs” refers to what the task would consider as its document (e.g., fact sentences for OpenBookQA). †Most questions in TriviaQA are answerable using only the first few hundred tokens in the document. . . . .	68
4.5	Hyper parameter values used in the Duet retrieval model. All layers are separated by ReLU activation units and a dropout layer with 0.5 probability. . . . .	76
4.6	Performance of different systems including the CSRQA model on our task. Hits@N scores reported in % , (p-value <0.0005). . . . .	76
4.7	Test set performance (Hits@3 in %) of ablation systems on questions with different candidate answer space sizes. . . . .	77
4.8	Performance of different systems including the CSRQA model on our task as measured using human judgements (Human Scores) and gold-reference data (Machine Scores). Hits@N scores reported in %. . . . .	78
4.9	The importance of question-specific entity embeddings generated using the QEA layer in CRQA . . . . .	79
4.10	Performance of CSRQA on the validation data reduces, as the size of candidate space (selected by CsQA) to be re-ranked increases. . . . .	80
4.11	Curriculum learning (CL) with different entity embedding schemes . . . .	81
5.1	Results of SPNET on the artificial spatial-questions dataset (t-test p-value < $10^{-33}$ for Hits@3) . . . . .	92
5.2	Performance of spatial-reasoning networks degrades in the presence of location-distractor sentences. . . . .	94
5.3	Performance of the BERT-BiLSTM CRF for tagging locations on a small set of 75 questions. . . . .	96
5.4	Distribution of questions with location-mention across train, dev & test sets.	97
5.5	Comparison of the joint Spatio-Textual model with baselines on questions that have location mentions (t-test p-value < 0.009) . . . . .	98
5.6	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 )	98

5.7	Comparison of Spatio-Textual CRQA (with and without (w/o) distance-aware question encoding) and CRQA on the full set . . . . .	98
5.8	Experiments on two subsets from the test-set: (i) Questions requiring Spatial-reasoning (ii) Questions with distractor-locations only. . . . .	100
5.9	Spatio-Textual CRQA: Classification of Errors . . . . .	100
5.10	Comparison with current state-of-the-art CSRQA on (i) Location Questions (ii) Full Task . . . . .	100
5.11	Hits@3 results on a blind-human study using 100 randomly selected questions from the test-set . . . . .	101
5.12	Comparison of re-ranking models operating on a reduced search space returned by CsQA on Location Questions (ii) Comparison of spatio-textual CSRQA+ with CSRQA and spatio-textual CSRQA on the full task. . .	102
6.1	Quality of extracted descriptive phrases on a devset . . . . .	111
6.2	Comparing clustering methods on development set . . . . .	116
6.3	User preference win-loss statistics for different clustering methods on both city and movie comparison task using the same IE system. Both EB G-pLSA and G-pLSA significantly outperform the baseline GMM model. EB G-pLSA has some edge over the G-pLSA model. Note: Ties have not been shown in the table. . . . .	116
A.1	City Wise - Knowledge Source Statistics . . . . .	132
A.2	City Wise Training Dataset Statistics . . . . .	133
A.3	City Wise Test Dataset Statistics . . . . .	134
A.4	City Wise Validation Dataset Statistics . . . . .	135
B.1	Templates used for generating the artificial dataset . . . . .	138
B.2	List of metonyms for each entity type in the artificial dataset . . . . .	139
B.3	Hyperparameter settings for experiments on the artificial-dataset . . . . .	139
B.4	Hyperparameters used for experiments on the end-task . . . . .	140
C.1	City Pairs used for comparing clustering algorithms . . . . .	142
C.2	Movie Pairs used for comparing clustering algorithms . . . . .	143
C.3	City Pairs used for evaluating summaries created by crowd source workers using the comparisons outputs from EB G-pLSA and by using full Wikipedia articles . . . . .	143