QA FOR TOURISM: ANSWERING RECOMMENDATION AND COMPARISON QUESTIONS

DANISH CONTRACTOR

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

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!

Danish Contractor

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 एंटिटी–खोज सुझाव प्रश्न के रूप में संदर्भित करते हैं, अर्थात, 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 एंटिटी का वर्णन करता है। हम एक सारणीबद्ध रूप में एंटिटी 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 ھی سفر کی منصوبہ بندی کے وقت وہ شہروں ' سیاحتی مقامات کے درمیان موازنہ سے متعلق سوالات بھی پوسٹ کرتے ہیں۔ان مقالہ جات میں ھم شعبہ سیاحت سے متعلق اس قسم کی تجاویز اورموازنہ کے جوابات دینے کے نئے کام کا جائزہ لینگے۔ھم اپنی توجہ تجاوزیز اور سوالات پر مرکوز کئے ہوئے ہیجو کسی چیز کی طالب ہیں۔ھم اسے ملٹی سینٹنس انٹیٹی سیکنگ تجاویز سوالات(MSRQs) قراردیتے ہیں مثلاً ایسے سوالات جو ایک یا ایک سے زائد جوابات چاھتے ہیں۔اس قسم کے انٹی ٹیزشعبہ سیاحت میں پوائنٹس آف انٹرسٹ کی(POIs) طرح ہوتے ہیں ' مثلاً ہوٹلس' رسٹورنٹ' سیاحتی مقامات ۔ھم انٹیٹی سیکنگ تجاوزیر اور سوالات کے دو ترتیب میں جواب دیتے ہیں انٹرمیڈیٹQA(i): تشریحات کیساتھQA(ii)انٹرمیڈیٹ تشریحات کے بغیر۔ھر ترتیب میں ھم ایک نیا مسئلہ وضع کرتے ہیں اور نئے ڈیٹاسیٹس بناتے ہیں جس سے امید ھے کہ ھمیآئندہ QA کی تحقیق میںمزید مدد ملے گی۔آخر میں ھم موازنہ کے سوالات کا جواب دینے سے متعلق مسئلہ کی بھی اسٹڈی کرتے ہیں۔ھم ٹیکسٹل کارپورا سے انٹیٹی موازنہ کو نکالنے کے کام کو وضع کرتے ہیں جس میں ھر ڈاکیومنٹ ایک وقت میں ایک انٹیٹی کو بیان کرتا ھے۔ھم انٹیٹی موازنہ کو جدولی شکل مینتیار کرتے ہیں جس میں اٹربیوٹ والیوفریسس' تجاوزیر کے فریسس اور دیگر وضاحتیں کلسٹرڈ اور لفظی اعتبار سے منظم ہوتی ھے۔جو براہ راست موزانہ کی اجازت دیتی ھے۔انٹیٹس کے بارے میں ھمارے جدولی خلاصے ' متوازن انفارمیشن جسکا موازنہ کیا جاتا ھے ۔ ھمارے استفادہ کنندوں کی تحقیق میں ھم نے یہ بھی پایا کہ استفادہ کنندے متوازن کلسٹرس کو زیادہ ترجیح دیتے ہیں۔

Contents

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; $\mathbb{P}$ = {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
ϕ	CRF Feature
ω	CRF Feature Weight
ρ_k	Weight associated with k^{th} constraint in Constraint Conditional Modeling (CCM)
$\mathcal{C}_k$	Violation score associated with the kth constraint in CCM
γ	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 ith location-mention
d_k	Distance of an entity from the kth 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
$\mathcal{S}_\mathcal{T}$	Textual Reasoning Score
ψ_T	Scaling weights for $\mathcal{S}_\mathcal{T}$
$\mathcal{S}_\mathcal{L}$	Spatial Reasoning Score
ψ_L	Scaling weights for $\mathcal{S}_\mathcal{L}$
$\mathcal{S}$	Entity Relevance Score
σ	Sigmoid function
$\bar{\rho}$	Spearman's rank coefficient
α, β	Weights for joint scoring in Spatio-Textual Reasoning
Θ	Parameters of a model
η	Regularization Weight for EB G-pLSA

Table 1: List of variables

List of Figures

List of Tables