RUTGERS
UNIVERSITY | NEW BRUNSWICK

Generating Fine-Grained Open Vocabulary Entity Type Descriptions

Rajarshi Bhowmik and Gerard de Melo

Introduction

- **Knowledge Graph**
  - Vast repository of structured facts

- **Why short textual description?**
  - Can succinctly characterize an entity and its type

- **Goal:** Generate succinct textual description from factual data

Motivating Problem

- Fixed inventory of ontological types (e.g. Person)

DBpedia Browse using Formats

About: Michael Jordan

An Entity of Type : person, from Named Graph : http://dbpedia.org,

Michael Jordan

American basketball player

WIKIDATA

Michael Jordan (Q41421)

American basketball player and businessman

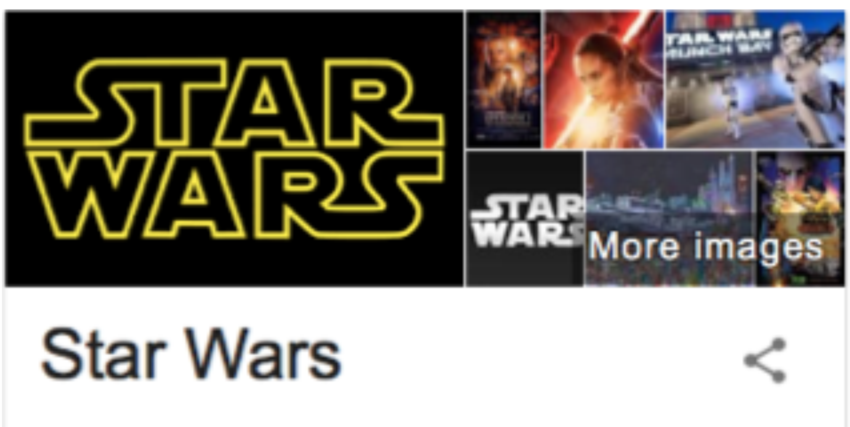
Motivating Problem

- Abstract ontological types can be misleading

About: Star Wars

An Entity of Type : sports team, from Named Graph : http://dbpedia.org,

- Missing short textual descriptions for many entities



Application: QA and IR

More Applications: Named Entity Disambiguation

RUTGERS

Desiderata

- ***Discerning most relevant facts***
  - Nationality and occupation for a person
    - E.g. “Swiss tennis player”, “American scientist”
  - Genre, regions and release year for a movie
    - E.g. “1942 American comedy film”
- Open vocabulary: applicable any kind of entity
- Generated text is ***coherent***, ***succinct*** and ***non-redundant***
- ***Sufficiently concise*** to be grasped at a single glance

RUTGERS

Key Contributions

- Dynamic memory-based generative model
  - jointly leverages ***fact embeddings*** + ***context of the generated sequence***
- Benchmark dataset
  - 10K entities with large variety of types
  - Sampled from Wikidata

Model Architecture

- 3 key modules:
  - Input Module
  - Dynamic Memory Module
  - Output Module

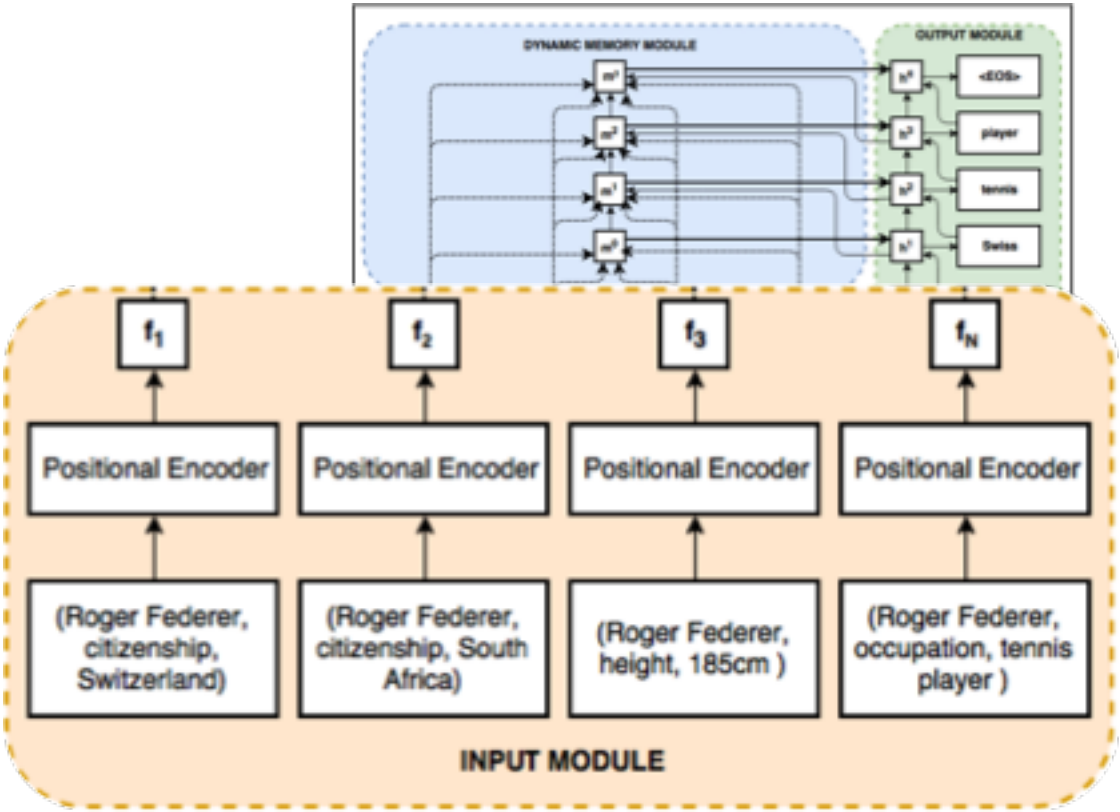
Input Module

- **Input**
  - set of N facts {f1, f2, …,fN}
- **Output**
  - concatenation of Fact Embeddings [**f1**, **f2**, …, **fN**]

- Learn Fact Embeddings using ***Word Embeddings*** + ***Positional Encoder***

- Positional Encoder:

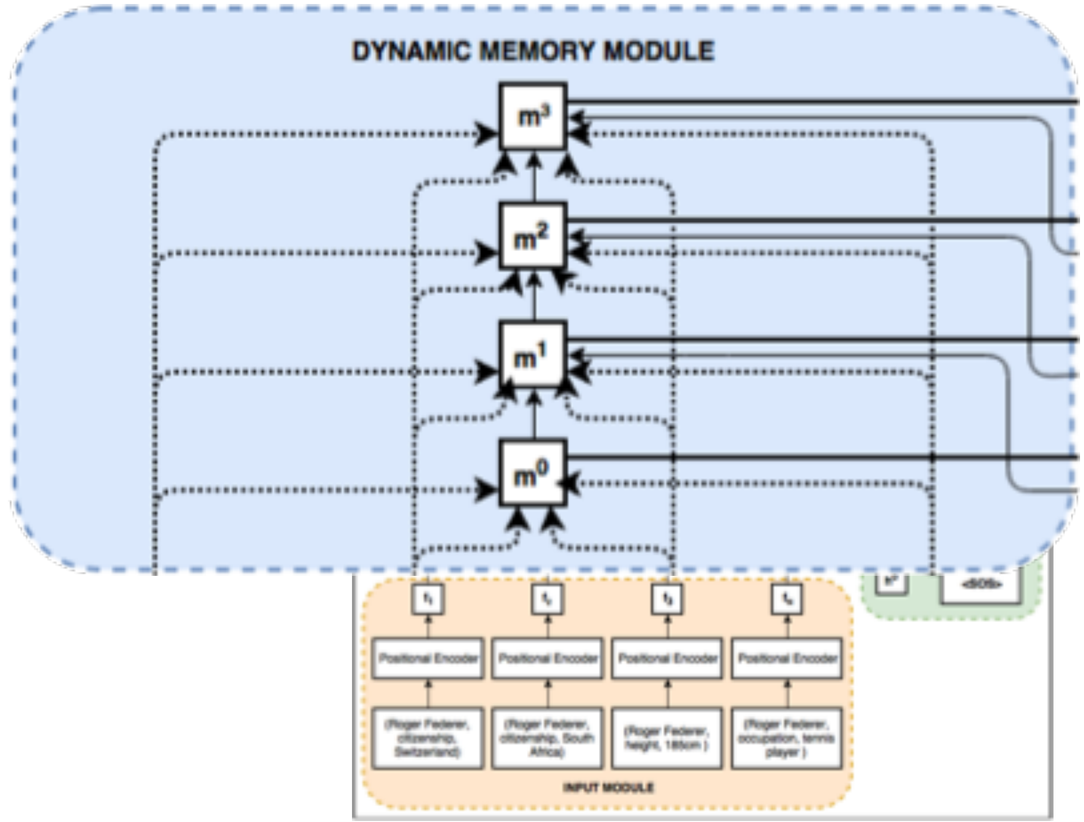
$$f_i = \sum_{j=1}^{J} l_j \circ w^i_j$$



Dynamic Memory Module

- Current context
  - Attention weighted sum of fact embeddings

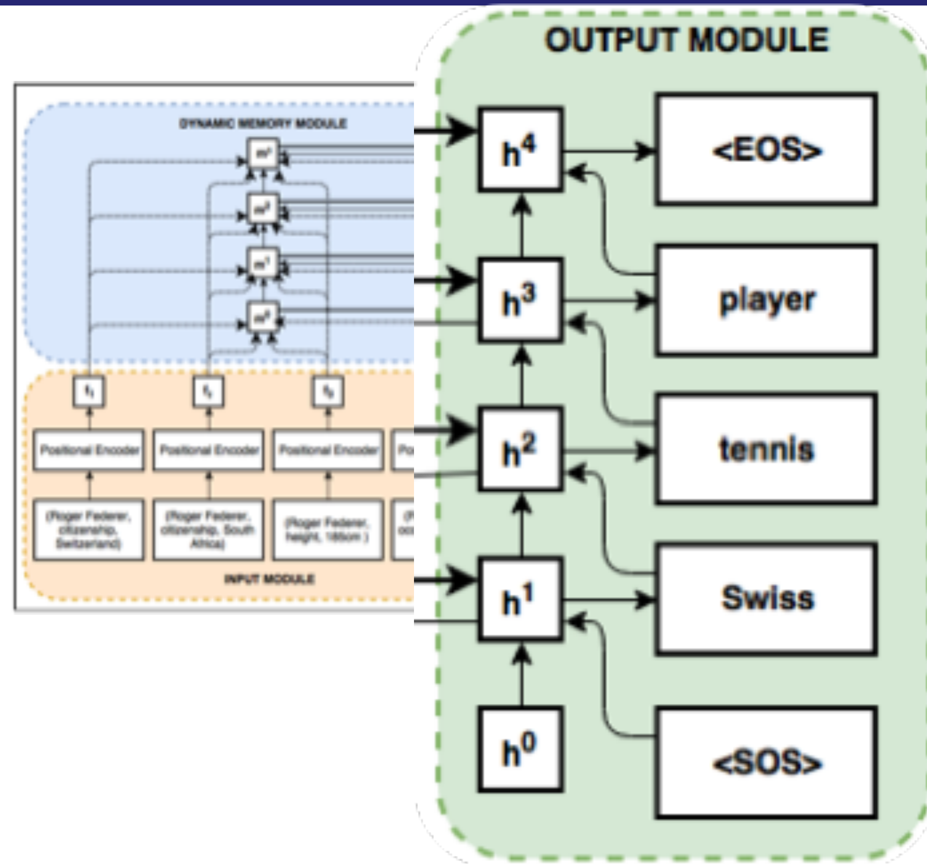
    $c^t = \sum_{i=1}^{N} a_i^t \boldsymbol{f_i}$
- Attentions weights depends on two factors:
  - How much information from a particular fact is used by the previous memory state
  - How much information of a particular fact is invoked in the current context of the output sequence
- Update memory state with
  - current context
  - previous memory state
  - current output context



Number of memory updates = Length of output sequence

Output Module

- Decode the current memory state to generate the next word
- Decoder GRU input:
  - current memory state m^t,
  - previous hidden state $h^{(t-1)}$
  - previous word $w^{(t-1)}$
    - During Training: ground truth
    - During evaluation: predicted word
- Concatenate output of GRU with the current context vector c^t
- Pass through a fully connected layer followed by a Softmax



RUTGERS

Evaluation: Benchmark Dataset Creation

- Sampled from Wikidata RDF dump and transformed to a suitable format
- Sampled 10K entities with a English description and at least 5 facts
- *fact* = (property name , property value).
- Transformed into a phrasal form by concatenating the words of the property name and its value
  - E.g. (*Roger Federer*, *occupation*, *tennis player*) → *'occupation tennis player'*

RUTGERS

Evaluation: Baselines

- Fact-to-sequence Encoder-Decoder Model
  - Sequence-to-sequence model (Sutskever et al.) is tweaked to work on the fact embeddings generated by positional encoder

- Fact-to-sequence Model with Attention Decoder
  - Decoder module uses an attention mechanism

- Static Memory
  - Ablation study *: No memory update* using the dynamic context of the output sequence

- Dynamic Memory Networks (DMN+)
  - Xiong et al.'s model with minor modifications
  - A question module gets a input question such as "*Who is Roger Federer?*" or "*What is Star Wars?*"

RUTGERS

Evaluation: Results

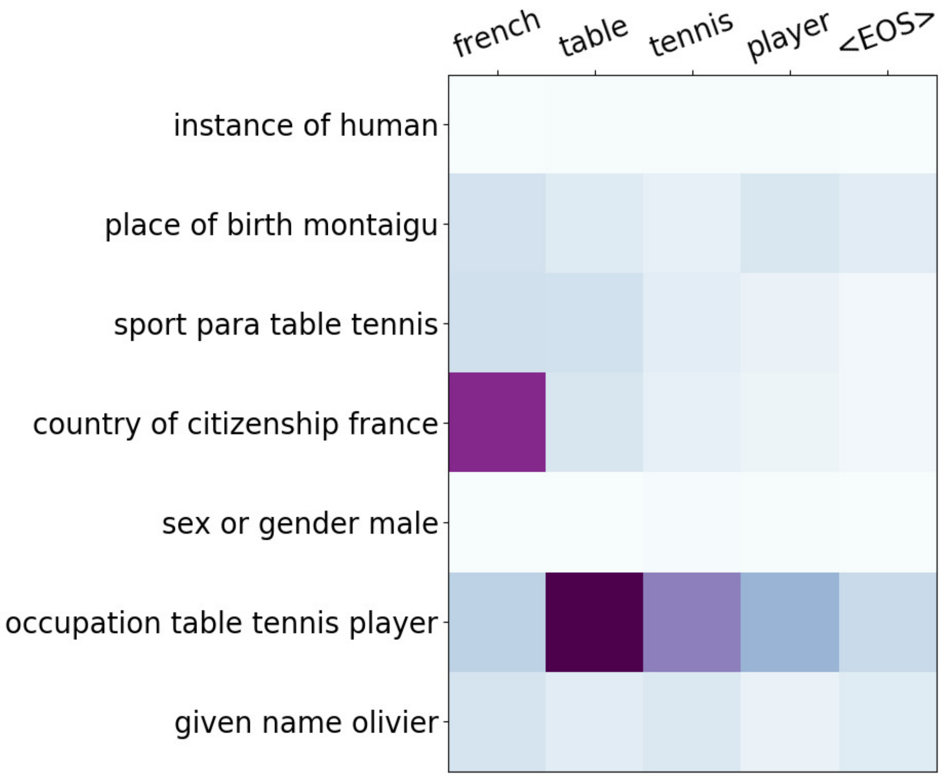
Model	B-1	B-2	B-3	B-4	ROUGE-L	METEOR	CIDEr
Facts-to-seq	0.404	0.324	0.274	0.242	0.433	0.214	1.627
Facts-to-seq w. Attention	0.491	0.414	0.366	0.335	0.512	0.257	2.207
Static Memory	0.374	0.298	0.255	0.223	0.383	0.185	1.328
DMN+	0.281	0.234	0.236	0.234	0.275	0.139	0.912
Our Model	**0.611**	**0.535**	**0.485**	**0.461**	**0.641**	**0.353**	**3.295**

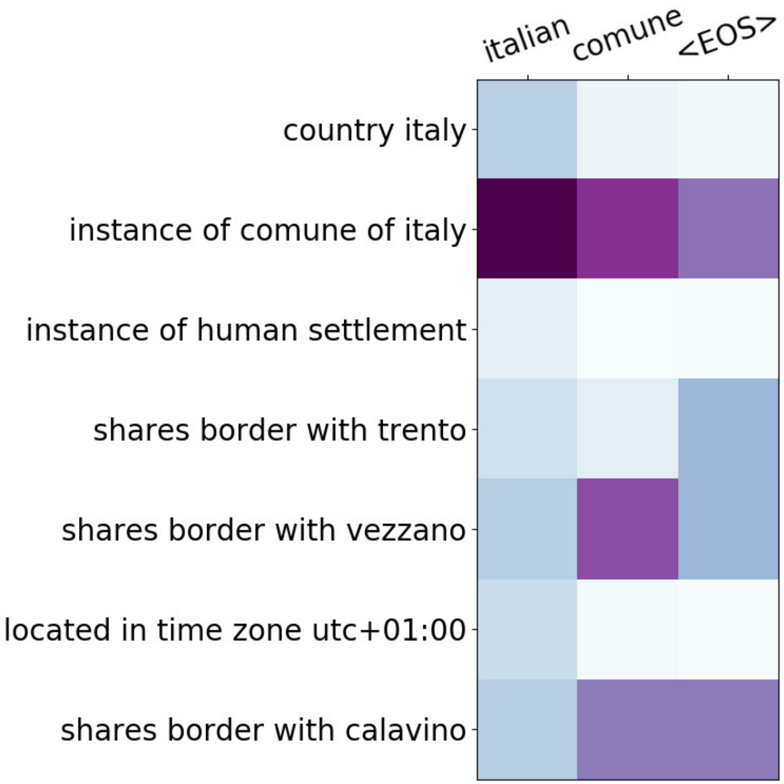
RUTGERS

Evaluation: Examples

	Wikidata Item	Ground Truth Description	Generated Description
Matches	Q669081	municipality in Austria	Municipality in Austria
	Q23588047	microbial protein found in Mycobacterium Abscessus	microbial protein found in Mycobacterium Abscessus
More specific	Q1865706	footballer	Finnish footballer
	Q19261036	number	natural number
More general	Q7815530	South Carolina politician	American politician
	Q4801958	2011 Hindi film	Indian film
Semantic drift	Q16164685	polo player	water polo player
	Q1434610	1928 film	filmmaker
Alternative	Q7364988	Dean of York	British academic
	Q1165984	cyclist	German bicycle racer

Evaluation: Attention Visualization





Conclusion

- Short textual descriptions facilitate instantaneous grasping of key information about entities and their types
- Discerning crucial facts and compressing it to a succinct description
- Dynamic memory-based generative architecture achieves this
- Introduced a benchmark dataset with 10K entities

Thank you!

https://github.com/kingsaint/Open-vocabulary-entity-type-description

Questions?