Tables as Semi-structured Knowledge for Question Answering

ACL'16

Sujay Kumar Jauhar[1] Peter D. Turney[2] Eduard Hovy[1]

[1]Carnegie Mellon University

[2]Allen Institute for Artificial Intelligence

Presented By : Dushyanta Dhyani

Outline

Outline

Problem Definition

Using tables for creating a Question Answering system.

This includes:

- Using tables for creating MCQ training Data.
- Building a QA model using tables

Outline

Primary Question : Why do we need tables?

Primary Question : Why do we need tables?

Subsequent Question: What are the alternatives to tables?

Primary Question : Why do we need tables?

Subsequent Question: What are the alternatives to tables?

- WEB

Primary Question : Why do we need tables?

Subsequent Question: What are the alternatives to tables?

- WEB
- Knowledge Bases

Primary Question : Why do we need tables?

Subsequent Question: What are the alternatives to tables?

- WEB
- Knowledge Bases
- Databases

Primary Question : Why do we need tables?

Subsequent Question: What are the alternatives to tables?

- WEB
- Knowledge Bases
- Databases

WEB

- Available in large volumes.
- Contain a large number of entities (And also continuously increasing at a rapid scale)
- Highly unstructured. Schema-less. Open domain.
- Thus difficult to reason with and interpret.
- Ensuring compositionality is difficult.
- Use Information Retrieval based methods.

Knowledge Bases

- Structured.
- Fixed Schema. Large number of entities and relations.
- Coverage is high.
- Compositionality is low.

Databases

- Structured. Fixed Schema.
- Few entities and relations.
- Ensure high compositionality.
- Extremely low coverage

TABLES!!!

TABLES!!!
Best of all the worlds

TABLES!!!
Best of all the worlds
Semi-structured

TABLES!!!
Best of all the worlds
Semi-structured
Can ensure compositionality

Motivation for carrying out the given work

- A few (77 out of 108) 4^{th} grade science exam questions from the Regent's dataset were manually annotated for alignment to tables.
- Built a QA System to solve the Aristo challenge.
- Rivaled the best solvers that AI2 had built till then.

Outline

Tables I

What are they?

- Semi-structured data

Phase Change		Initial State		Final State		Form of Energy Transfer
Melting	causes a	solid	to change into a	liquid	by	adding heat
Vaporization	causes a	liquid	to change into a	gas	by	adding heat
Condensation	causes a	gas	to change into a	liquid	by	removing heat
Sublimation	causes a	solid	to change into a	gas	by	adding heat

- Contain general knowledge data.
- Cells contain free form text. Can be used independently as raw text.
- Thus also help in comparing the proposed model to an information retrieval based model.

Tables II

What are they?

- However, each row exhibits a well-defined recurring filler pattern.
- Help in providing rich alignment annotations.
- **Content Columns** : Columns with explicitly specified headers. These columns contain concepts, entities, processes, etc.
- **filler columns** : Columns with no headers. Contain a recurring pattern.

- The topics for creating these tables were determined from the Training set of 2 evaluation datasets : **Regents** & **Monarch**
- **Regents** Dataset:
  - Public Dataset
  - Single table containing 108 questions.
  - Questions revolve around 4^{th} grade science.
- **Monarch** Dataset:
  Unreleased Dataset also related to 4^{th} grade science.

Tables

Tables for this task?

- AI2's Aristo Tablestore
- 65 hand crafted tables organized by topics.
- Table Topics - Bounded & Unbounded
- Total : 3851 facts (Row=fact)

Table Name	**Facts**	**Table Name**	**Facts**
Orbital Event Daylight Hours	4	Country Hemispheres	267
Phase Transitions	6	Device Energy Conversion	65
Average Weights of Animals	1225	Wordnet Definitions	2467

Outline

TabMCQ

Crowdsourcing!!!

- Amazon Mechanical Turk!!
- Structural constraints are imposed while creating questions.
- In case of insufficient choices for an answer, Turkers provide answers on their own.

TabMCQ

Crowdsourcing!!!

Phase Change		Initial State		Final State		Form of Energy Transfer
Melting	causes a	solid	to change into a	liquid	by	adding heat
Vaporization	causes a	liquid	to change into a	gas	by	adding heat
Sublimation	causes a	solid	to change into a	gas	by	adding heat
Freezing	causes a	liquid	to change into a	solid	by	removing heat
Deposition	causes a	gas	to change into a	solid	by	removing heat
Condensation	causes a	gas	to change into a	liquid	by	removing heat

- Answer
- Cells that could be a part of the question
- Distractors / Alternate Answers

TabMCQ

How efficient is it?

Task	Avg. Time (s)	$/hour	% Reject
Rewrite	345	2.61	48
Paraphrase	662	1.36	49
Add choice	291	2.47	24
Write new	187	**5.78**	38
TabMCQ	**72**	5.00	**2**

Outline

- For generating MCQ data, given a cell in a table, we generated question from cell's row and candidate answers from the cell's column.
- Now the task at hand is reversed.
- Given question-answer pairs, find a cell that best confirms the assertion.

- Log-linear Model that assigns a score to every cell in every table according to their relevance to each question-answer pair
- Formally, Given :
  - $\mathbf{Q} = \{q_1, q_2, ...q_N\}$ - set of Questions
  - $A_n = \{a_n^1, a_n^2, ...a_n^k\}$ for a given question q_n
  - $\mathbf{T} = \{T_1, T_2, ...T_M\}$ - Set of Tables

$$log\, p(t_m^{ij}|q_n, a_n^k; A_n, T) = \sum_d \lambda_d f_d(q_n, a_n^k, t_m^{ij}; A_n, T) - logZ$$

where,

λ_d is a set of parameters to be learned.

$f_d(....)$ is a set of features

$Z = \sum_{m,i,j} exp\left(\sum_d \lambda_d f_d(q_n, a_n^k, t_m^{ij}; A_n, T)\right)$ is the partition function.

$t_m^{ij} = Cell(i, j)$ in the m^{th} table.

- We try to assess the significance/salience of each cell for a given **Q-A** pair.
- Relevant cells , if present, assert to the hypothetical claim/fact made by a given QA pair.
- During Inference, we decide upon the answer choice that gets the maximum score out of all the rows i.e.

$$a_n^* = \arg\max_{a_n^k} \max_{m,i} \sum_j \sum_d \lambda_d f_d(q_n, a_n^k, t_m^{ij}; A_n, T)$$

- For training, apart from the features, we also need predictor values.
- Since we are dealing with probability of cell relevance, the predictor value should numerically quantify the alignment of rows to questions and columns to answer choices.
- Authors selected the following methodology from a few (tested) other scoring heuristics.
  - For a **correct** QA pair,
    - Assign a score of **1.0** to cell that exactly answers the question.
    - If a cell does not answer the question, but is used in the construction of the question, assign a score of **0.5**
    - Otherwise assign a score of **0.0**
  - For an **incorrect** QA pair:
    - With a probability of 1% , assign a score of 0.1 to random cells from all the (other) tables that have no alignment to the given QA pair.

Cross-Entropy Loss Function

$$L(\vec{\lambda}) = \sum_{\substack{q_n \\ a_n^k \in A_n}} \sum_{m,i,j} p(t_m^{*ij} | q_n, a_n^k; T).log\, p(t_m^{ij} | q_n, a_n^k; A_n, T)$$

where $p(t_m^{*ij} | q_n, a_n^k; T)$ is the normalized probability of the true alignment scores

Adaptive Gradient Descent (AdaGrad) is used to minimize the above loss.

FRETS

Rich Features?

- Several features are used at various granularity levels : Table , Row , Column , Cell
- Certain features are supplemented with their soft matching variants.
- Features that are assigned high weights (λ_d) during training are then used to form a compact FRETS model.

Level	Feature	Description	Intuition
Table	Table score	Ratio of words in **t** to **q+a**	Topical consistency
	†TF-IDF table score	Same but TF-IDF weights	Topical consistency
Row	Row-question score	Ratio of words in **r** to **q**	Question align
	Row-question w/o focus score	Ratio of words in **r** to $\mathbf{q}$-$(\mathbf{a_f}+\mathbf{q_f})$	Question align
	Header-question score	Ratio of words in **h** to **q**	Prototype align
Column	Column overlap	Ratio of elements in **c** and **A**	Choices align
	Header answer-type match	Ratio of words in $\mathbf{c_h}$ to $\mathbf{a_f}$	Choices hypernym align
	Header question-type match	Ratio of words in $\mathbf{c_h}$ to $\mathbf{q_f}$	Question hypernym align
Cell	†Cell salience	Salience of **s** to **q+a**	QA hypothesis assert
	†Cell answer-type entailment	Entailment score between **s** and $\mathbf{a_f}$	Hypernym-hyponym align
	Cell answer-type similarity	Avg. vector sim between **s** and $\mathbf{a_f}$	Hypernym-hyponym sim.

FRETS

Feature Explanation - TF-IDF weighting

- TF-IDF scores are computed for all words in all the tables.
- Each table is treated as a unique document.
- *"...At run-time we discount scores by table length as well as length of the QA pair under consideration to avoid disproportionately assigning high scores to large tables or long MCQs."*

FRETS

Feature Explanation - Focus Words

- **Question & Answer Focus**: Parse questions to find question type & desired answer type.
  e.g. Given "**What form of energy is required to convert water from a liquid to a gas?**"
  Answer Type: *Form of Energy*
- Based on question patterns in the data, a rule-based parser (using a set of hand-coded regular expressions) is used to find answer-types from queries.
- This parser is designed such that it produces answer types only in high confidence situations.
- Similar operation is performed for questions.

- **Salience** for a pair of strings is evaluated by computing Point-wise Mutual Information (PMI) statistics between this pair from a large corpus.
- The higher the salience score, the higher the relevance of the cell for the QA hypothesis.
- **Entailment** refers to the confidence in the truthfulness of one string, given that another string is **true**.
- Features used for evaluating entailment - Overlap, Paraphrase probability, lexical entailment likelihood and ontological relatedness.

FRETS

Features - Soft Matching & Compactness

Level	Feature	S-Var	Cmpct
Table	Table score	◊	
	†TF-IDF table score	◊	●
Row	Row-question score	◊	●
	Row-question w/o focus score	◊	
	Header-question score	◊	
Column	Column overlap	◊	●
	Header answer-type match	◊	●
	Header question-type match	◊	
Cell	†Cell salience	◊	●
	†Cell answer-type entailment		●
	Cell answer-type similarity		

Features - Soft Matching

- Most of the features used, primarily follow a bag of words based model.
- Thus a hard overlap feature would define a score between two bag of words S_1 & S_2 as $|S_1 \cap S_2|/|S_1|$
- However, it's quite possible that a new QA pair might not contain the terms in the vocabulary of the model.
- Thus a soft-matching counterpart of the above features is proposed.

FRETS II

Features - Soft Matching

- Similarity between words in an embedding space is used as an alternate to $|S_1 \cap S_2|$
- Thus the corresponding soft overlap score would be:

$$\frac{1}{|S_1|} \sum_{w_i \in S_1} \max_{w_j \in S_2} sim(\vec{w_i}, \vec{w_j})$$

- Word embeddings are obtained by training on 300 million words of the **WMT-2011** shared task and were improved by retrofitting [1] them to PPDB - paraphrase database [2]

Outline

- **Evaluation Dataset** :
  - **Regents** - publicly available ; **129** MCQ's
  - **Monarch** - Unreleased ; **250** MCQ's
  - **Elementary School Science Questions (ESSQ)** - Public dataset ; **855** MCQ's
- Since the training tables of **Regents** & **Monarch** were used in the construction of the Aristo Tablestore, only the testing set was used for evaluation. All tables from ESSQ were used.

- **Information Retrieval Method**
  - Uses Lucene search engine.
  - Table structure is ignored. The rows are used as simple text.
  - Top results from Lucene are then used to rank the different answer choices.
- **Markov Logic Networks** (MLN) [3]:
  - Highly structured models.
  - Results directly cited from [4].
  - Thus results only available for **Regent's** dataset.

3 different combinations of training data were used:

- Aristo tables only constructed for **Regent's** dataset. Total = **40** tables.
- Aristo tables only constructed for **Monarch** dataset. Total = **25** tables.
- All the above. Total = **65** tables.

For the Lucene baseline, they also experimented with the **Waterloo corpus** that contain $5 * 10^1 0$ words

FRETS
Results

Model	Data	Regents Test	Monarch Test	ESSQ
Lucene	Regents Tables	37.5	32.6	36.9
	Monarch Tables	28.4	27.3	27.7
	Regents+Monarch Tables	34.8	35.3	37.3
	Waterloo Corpus	55.4	51.8	54.4
MLN (Khot et al., 2015)	-	47.5	-	-
FRETS (Compact)	Regents Tables	**60.7**	47.2	51.0
	Monarch Tables	56.0	45.6	48.4
	Regents+Monarch Tables	59.9	47.6	50.7
FRETS	Regents Tables	59.1	**52.8**	54.4
	Monarch Tables	52.9	49.8	49.5
	Regents+Monarch Tables	59.1	52.4	**54.9**

- Without the **Waterloo Corpus**, the Lucene baseline has scores far less than FRETS.
- Thus FRETS is able to outperform an unstructured model given small amount of data.
- In certain cases, FRETS performs significantly better than MLN, a model that is highly structured, with a much complex data formalism.
- In the **Lucene** baseline, the performance decreases in case of **Monarch** or **Regents+Monarch** tables as training data. However, that is not the case with FRETS or not very significant in case of FRETS(Compact). This shows the addition of tables (presumably not useful) does not affect the performance of FRETS.

FRETS
Ablation Study

A set of features were removed and the model's performance was analyzed.

Model	REG	MON	ESSQ
FRETS	59.1	**52.4**	**54.9**
w/o tab features	59.1	47.6	52.8
w/o row features	49.0	40.4	44.3
w/o col features	59.9	47.2	53.1
w/o cell features	25.7	25.0	24.9
w/o ♢ features	**62.2**	47.5	53.3

Outline

Conclusion

- Tables can be used as knowledge bases for QA.
- A connected framework is proposed for both dataset generation and MCQ solving.
- Trade off between structure and reasoning ability of tables is efficiently handled.
- A large dataset of more than 9000 MCQ questions is publicly released.

References I

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith.
Retrofitting word vectors to semantic lexicons.
arXiv preprint arXiv:1411.4166, 2014.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch.
Ppdb: The paraphrase database.
2013.

Tushar Khot, Niranjan Balasubramanian, Eric Gribkoff, Ashish Sabharwal, Peter Clark, and Oren Etzioni.
Exploring markov logic networks for question answering.
In *Proceedings of EMNLP*, volume 5, 2015.

References II

Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Turney, and Daniel Khashabi.
Combining retrieval, statistics, and inference to answer elementary science questions.
30th AAAI, 2016.