Making Sense of Big Data Using Question Answering

Alvaro Morales, Sue Felshin, Boris Katz

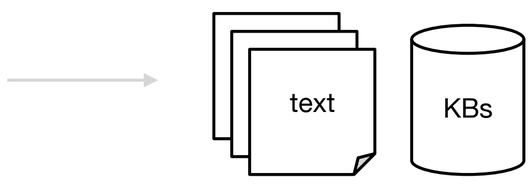
CSAIL InfoLab

From processing to understanding

By focusing on parts of the data that are interesting to a human user, **question answering** systems can reduce massive amounts of information into a relevant answer.

We can provide the user with high-precision access to just the right information by **exposing relations and constraints in language**.

```
Which river does the
Brooklyn Bridge cross?
```



Recent advances enable us to scale up and **process** massive amount of text. We need better systems to **understand** natural language in **highly varied** repositories like the Web.

Efforts in information extraction need to be enhanced with natural language understanding.

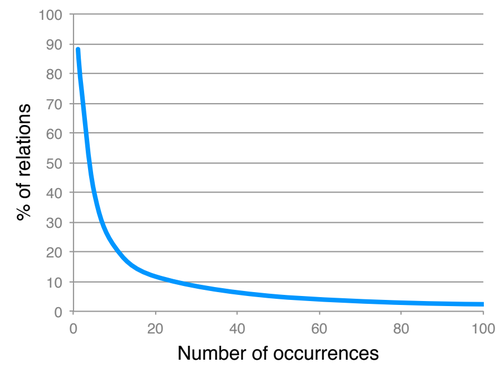
Sparseness in knowledge bases

In recent years, knowledge bases (KBs) constructed automatically from unstructured and semi-structured data have emerged.

Systems have been getting better in completeness and precision, but **sparsity** in relation phrases is still a big problem.

Case Study: Reverb [1]

- Extractors ran on ClueWeb09, a 25 TB dataset of 1B web pages
- 6B extractions. 15M extractions with confidence > 90%
- Out of 700K normalized relation phrases, only 5% have more than 50 occurrences



Methods need to work **across** knowledge bases.

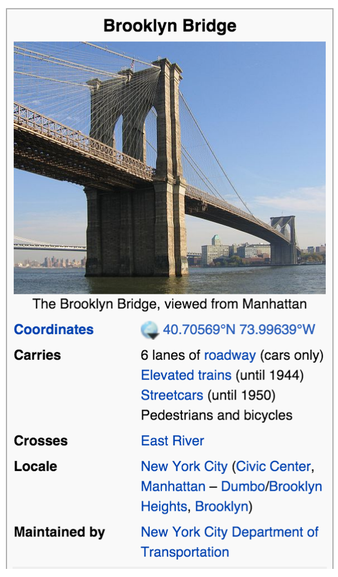
[1] Fader, Anthony, Stephen Soderland, and Oren Etzioni. "Identifying relations for open information extraction." Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011.

Ambiguity in extractions

Knowledge bases contain triples of subject-relation-object

```
<Falcone> <is incarcerated in> <Arkham>
```

Triples can be generated from **semi-structured sources** like Wikipedia infoboxes




```
<Brooklyn Bridge> <crosses> <East River>
<Brooklyn Bridge> <coordinates> <40°N,73°W>
```

Triples may suffer from **syntactic ambiguity**

To increase the precision and usefulness of extracted triples, we find **semantic data types** associated with the arguments

```
<any-bridge> <crosses> <any-river>
```

Search for matching sentences in free text. Use dependency parse of sentence to disambiguate

Paraphrasing and subsuming relations

It is **non-trivial** to understand semantics of a relation and know alternative ways to ask about it.

went to school	is a county located in
attended college	was located on
alma mater	was originally located at
graduated from	is in

A relation phrase may **subsume** another relation phrase. Define subsumption from a dependency parse.

```
was located on, was originally located at
was <ADV>* located <IN>
```

Use subsumptions and semantic data types to **cluster** similar relation phrases.

Reducing sparsity (while preserving expressivity) makes a KB **more accessible** to a question answering system.

Generalized annotations

We can make knowledge bases more useful and accessible by generating **annotations** generalized over semantic data types

Annotations are used by the START system, developed at the InfoLab, to answer questions

Relation	Annotation	Sample Question
`institutions`	`any-scientist` teaches at `any-university`	Where does Marvin Minsky teach?
`alma-mater`	`any-officeholder` graduated from `any-university`	Where did Barack Obama graduate from?
`current-team`	`any-nba-player` plays with the `any-nba-team`	Who does LeBron James play with?
`apprehended`	`any-murderer` was captured on `any-date`	When did Ted Bundy get captured?
`awards`	`any-scientist` has been given many awards	What awards were given to Stephen Hawking?

With START's capabilities, a single annotation can match multiple syntactic forms of a question, and can answer questions about all entities of the same type

Applications to question answering

Question answering is an **effective interface** to open up a massive dataset to exploration by users

Natural language understanding can make big data **more useful and interactive**

It is easier to **reason** over a more structured knowledge base, and perform **inference** over relations

```
study  ——→  attend university
```

With a more robust understanding of the knowledge base, it is possible to **explain** the system's decisions

```
==> Where did Bill Clinton study?
Georgetown University

Explanation: I looked for a university that Bill Clinton
attended. I matched this answer to your question with high
confidence.
```