Do it your own (DIY) Jeopardy Question Answering System

André Freitas and Edward Curry

Digital Enterprise Research Institute (DERI)
National University of Ireland, Galway

1 Motivation

The evolution and maturity of semantic technologies techniques and frameworks are bringing functionalities which were once considered academic or prototypical into real-life applications. Products such as *IBM Watson* [1] and *Siri* are examples of applications which are heavily leveraged on state-of-the-art semantic technologies. These systems provide a synthesis of the functionalities which are available for general applications today such as: natural language search and queries over large-scale data, semantic flexibility and integration between structured and unstructured resources. The success of these projects in demonstrating the potential of existing technologies lies on the fact that they bring into a single system approaches from Natural Language Processing (NLP), Semantic Web (SW), Information Retrieval (IR) and Databases.

This work demonstrates *Treo*, a framework which converges elements from NLP, IR, SW and Databases, to create a semantic search engine and question answering (QA) system for heterogeneous data. *Jeopardy* and *Question Answering* queries over open domain structured and unstructured data are used to demonstrate the approach. In this work, *Treo* is extended to cope with unstructured text in addition to structured data. The setup of the framework is done in 3 steps and can be adapted to other datasets in a simple DIY process.

2 Treo: Querying structured & unstructured data

Treo supports free natural language queries over both structured and unstructured data. To enable *semantic flexibility* and *vocabulary independency* in the query process, a principled *distributional-compositional semantic model* is used to build a distributional structured vector space model ($\tau - Space$) [2]. *Distributional semantics* focuses on the *automatic construction* of a semantic model based on the statistical distribution of co-located words in large-scale corpora. The distributional semantics component of the model, supports a semantic approximation between query and dataset terms: operations in the $\tau - Space$ are mapped to *semantic relatedness* operations using the distributional model as a commonsense knowledge base [2]. The automatic creation of distributional semantic models supports the *transportability* of the approach to other datasets and languages, not requiring the manual creation effort of ontologies (Treo does not rely on ontology-based reasoning for semantic approximation).

In addition to queries over structured data, this work extends the query mechanism for searching entities in unstructured text. Both structured and unstructured data are linked in an entity-centric semantic index (Figure 1 (B)). The elements of the query processing approach are depicted in Figure 1 (A).

Two different query processing strategies are used:

- Query processing over structured data: In the *query pre-processing* phase, the natural language query is analyzed by the *Interpreter* component, where a set of *query triple patterns and features* are detected in the user query. The second phase consists of the *vocabulary independent query processing approach* which defines a sequence of search and data transformation operations over the structured data graph embedded in the $\tau - Space$ [2], targeting the maximization of the semantic matching with the query. The *Query Planner* generates the sequence of *semantic search, navigation and transformation operations* over the graph data, which defines the *query processing plan*, based on a set of *query features* which are determined in the pre-processing phase. The third phase consists in the execution of the query processing plan operations over the $\tau - Space$ index.

- Query processing over structured & unstructured data: In case the query is not addressed by the available structured data, the query can be processed against both structured data and unstructured text in the entity-centric index. The query pre-processing approach for this query type consists on the detection of the *query focus* by the application of POS Tag based rules and by the detection and resolution of *named entities* in the query. The *query plan* consists of the composition of keyword-search operations over the text segments associated with entities, distributional search operations over structured data, and keyword search over associated entities. A *ranking function* weights the results of all operations, also taking into account the *cardinality* for each entity (number of associated entities, facts and text segments). The initial top-20 entity results are re-ranked based on the computation of the *distributional semantic relatedness scores* between the *query focus phrase* and the associated *entity types.*

3 DIY Setup Process

The setup of the Treo platform for a new dataset consists in the creation of a *semantic index* for both structured and unstructured data, which requires three steps:

1. *Construction of the distributional semantic model:* Consists on the use of a large-scale reference corpora to build the distributional semantic reference model [2]. In this demonstration Wikipedia 2006 is used as the reference corpus and Explicit Semantic Analysis (ESA) is the distributional semantic model.
2. *Semantic indexing of structured data:* Consists in the indexing of structured data using the distributional semantic reference model [2]. The framework

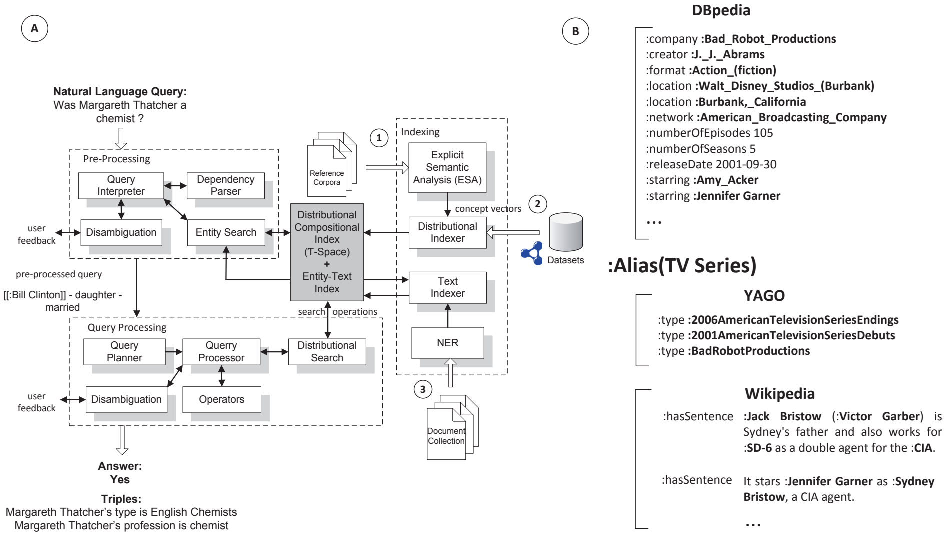

A
B
Natural Language Query:
Was Margareth Thatcher a chemist ?
Pre-Processing
Query Interpreter
Dependency Parser
user feedback
Disambiguation
Entity Search
pre-processed query
[[:Bill Clinton]] - daughter - married
Query Processing
Query Planner
Query Processor
Distributional Search
user feedback
Disambiguation
Operators
Answer:
Yes
Triples:
Margareth Thatcher's type is English Chemists
Margareth Thatcher's profession is chemist
Reference Corpora
1
Indexing
Explicit Semantic Analysis (ESA)
concept vectors
2
Distributional Compositional Index (T-Space) + Entity-Text Index
Distributional Indexer
Datasets
Text Indexer
search operations
NER
3
Document Collection
DBpedia
:company :Bad_Robot_Productions
:creator :J._J._Abrams
:format :Action_(fiction)
:location :Walt_Disney_Studios_(Burbank)
:location :Burbank,_California
:network :American_Broadcasting_Company
:numberOfEpisodes 105
:numberOfSeasons 5
:releaseDate 2001-09-30
:starring :Amy_Acker
:starring :Jennifer Garner
...
:Alias(TV Series)
YAGO
:type :2006AmericanTelevisionSeriesEndings
:type :2001AmericanTelevisionSeriesDebuts
:type :BadRobotProductions
Wikipedia
:hasSentence :Jack Bristow (:Victor Garber) is Sydney's father and also works for :SD-6 as a double agent for the :CIA.
:hasSentence It stars :Jennifer Garner as :Sydney Bristow, a CIA agent.
...

Fig. 1: (A) Semantic indexing and query processing architecture. (B) Entity-centric representation of structured and unstructured data.

takes as input data any dataset following an Entity-Attribute-Value (EAV) format. DBpedia 3.7 and YAGO are used as the demonstration datasets.

3. *Unstructured data entity-centric indexing:* This step takes as input a text collection, recognizes the named entities based on the structured data previously indexed, aligning it with the indexed structured data. The demonstration uses Wikipedia 2013 as the test collection.

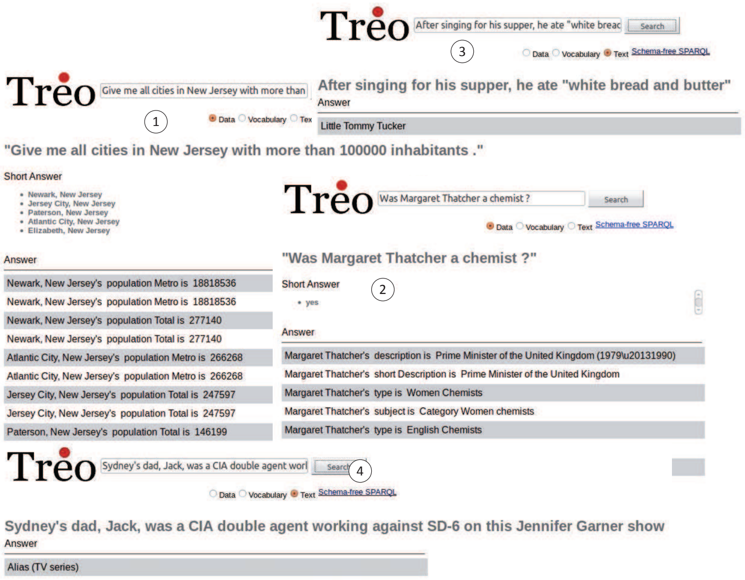
The steps are executed by calling one script, which takes as input the three types of resources (reference corpora, structured datasets and unstructured texts). After the setup, natural language queries can be executed against the structured and unstructured data indexes. Figure 1 shows the components of the Treo architecture (A) and an example of the entity-centric linking between structured and unstructured data (B).

4 Demonstration

The system is demonstrated over the open-domain *DBpedia 3.7/YAGO* RDF datasets and Wikipedia 2013 text data. The RDF datasets consist of 128,071,259 triples (17GB) loaded into the Treo index for structured data. A set of natural language queries from the *Jeopardy challenge*[1] and from the *Question Answering over Linked Data* challenge[2] are used to demonstrate the system. In the demonstration, users input free natural language queries and the system returns two

[1] http://j-archive.com/

[2] QALD-1, http://www.sc.cit-ec.uni-bielefeld.de/qald-1, 2011



Treo [Give me all cities in New Jersey with more than]

● Data ○ Vocabulary ○ Tex

"Give me all cities in New Jersey with more than 100000 inhabitants ."

Short Answer

- Newark, New Jersey
- Jersey City, New Jersey
- Paterson, New Jersey
- Atlantic City, New Jersey
- Elizabeth, New Jersey

Answer

Newark, New Jersey's population Metro is 18818536
Newark, New Jersey's population Metro is 18818536
Newark, New Jersey's population Total is 277140
Newark, New Jersey's population Total is 277140
Atlantic City, New Jersey's population Metro is 266268
Atlantic City, New Jersey's population Metro is 266268
Jersey City, New Jersey's population Total is 247597
Jersey City, New Jersey's population Total is 247597
Paterson, New Jersey's population Total is 146199

Treo [Was Margaret Thatcher a chemist ?] Search

● Data ○ Vocabulary ○ Text Schema-free SPARQL

"Was Margaret Thatcher a chemist ?"

Short Answer

- yes

Answer

Margaret Thatcher's description is Prime Minister of the United Kingdom (1979\u20131990)
Margaret Thatcher's short Description is Prime Minister of the United Kingdom
Margaret Thatcher's type is Women Chemists
Margaret Thatcher's subject is Category Women chemists
Margaret Thatcher's type is English Chemists

Treo [After singing for his supper, he ate "white breac] Search

○ Data ○ Vocabulary ● Text Schema-free SPARQL

After singing for his supper, he ate "white bread and butter"

Answer

Little Tommy Tucker

Treo [Sydney's dad, Jack, was a CIA double agent worl] Search

○ Data ○ Vocabulary ● Text Schema-free SPARQL

Sydney's dad, Jack, was a CIA double agent working against SD-6 on this Jennifer Garner show

Answer

Alias (TV series)

Fig. 2: Example queries: (1,2) Queries over structured data (3,4) Jeopardy queries over structured and unstructured data.

types of results: (i) a list of highly related triples or (ii) post-processed results, depending on the query type.

Figure 2 (2) shows the output of a query over the structured data index for the query *'Was Margaret Thatcher a chemist?'*. In addition to the post-processed answer, which provides a direct (QA-style) answer for the query, the mechanism shows the justification for the answer with the supporting triples. Figure 2 (1) shows a query over structured data with a complex query plan (*'Which cities in New Jersey have more than 10000 inhabitants?'*). Figure 2 (3) and (4) show examples of Jeopardy queries, which typically provide a natural language description of a named entity or concept (for example: *'Sydney's dad, Jack, was a CIA double agent working against SD-6 on this Jennifer Garner show'*). Further examples can be found online[3].

Acknowledgments. This work was funded by SFI Ireland (SFI/08/CE/I1380).

References

1. D. Ferrucci et al., Building Watson: An Overview of the DeepQA Project, AI Magazine, 2010.
2. A. Freitas, E. Curry, J. G. Oliveira, S. O'Riain, A Distributional Structured Semantic Space for Querying RDF Graph Data. International Journal of Semantic Computing (IJSC), 2012.

[3] http://treo.deri.ie/ISWC2013Demo