
Semantic Indexing

Anant Bhardwaj
Ekaterina Ob'yedkova
Ravi Netravali

ANANTB@CSAIL.MIT.EDU
KATYA_@MIT.EDU
RAVINET@MIT.EDU

1. Abstract

Majority of information retrieval and web search use the inverted index as the backbone for their keyword based search. Unfortunately, inverted index fail to capture the semantic of the language and thus keyword based approach can only go as far as giving the relevant hits. We propose semantic indexing scheme and a new search interface that will not only augment keyword based search approach but would also answer questions that are posed to it in natural language.

The indexer would do a semantic encoding of english text corpus to build its index store. As new data comes in, the index store would get updated. For querying, we would allow a web interface that can take natural language query and would search the index store to find an answer to the query. We would also provide APIs for building the index and a SQL/python interface for searching the index.

The semantic encoding in our proposal would involve parsing of the english sentences which would improve upon the work done by Matthew Fay and Jodyann Coley. Some of the improvements we propose are:

1. Probabilistic Parsing: Our system would incorporate a probabilistic sentence parser which would allow the system to intelligently identify the part-of-speech of previously unknown words and appropriately add them to the index store.
2. Multiple adverbs and prepositional phrases: The current implementation only allows a single occurrence of location, adverb, and instrument phrases. Our system would allow sentences such as "John slept under the tree in the park" to be understood, and the question "Where did John sleep" would return "under the tree in the park".
3. Noun Phrases with relative classes, and with time: Show me all the widgets that were selected last Tuesday.
4. Handling complex negation (negation of x AND Y kind of phrases)
5. Intersective vs. Non-intersective adjectives
6. Complex causality handling: The current implementation only allows a primitive causality of 'if then' kind. Expanding to to more complex causality mapping would allow us handle complex causality.
7. Handling constraints around conjunctions and disjunctions.