Cognitive Computation Group
University of Illinois at Urbana-Champaign

New Search Paradigms for Meaning-based Information Retrieval

Quang Xuan Do, V.G.Vinod Vydiswaran, and Dan Roth

Homeland Security
MIAS
Multimodal Information Access and Synthesis

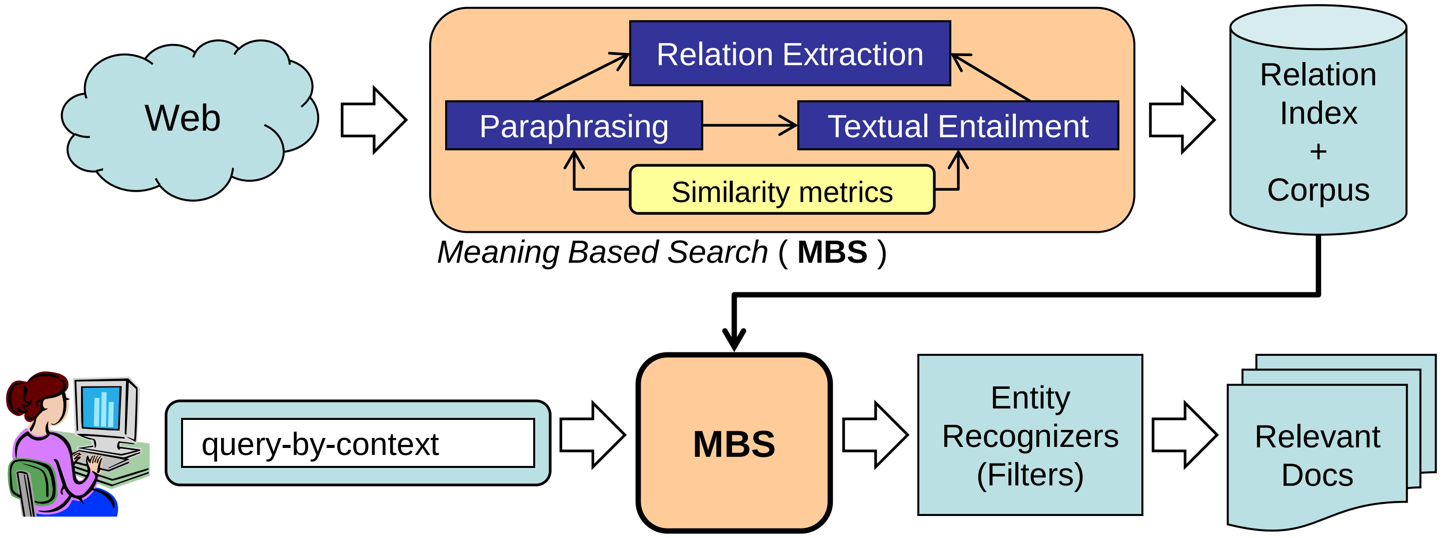
The fundamental operation used to access unstructured information is that of **Search**. Traditional keyword-based Search has been popularized by commercial search engines like Google and Yahoo!, where retrieval is heavily dependent on **keywords** and their frequency in target documents. However, such techniques often fail to capture the informational needs of the users, resulting in failure to retrieve essential information even when it is available. The goal of this project is to develop natural language processing capabilities and associated search protocols that improve search capabilities. Specifically, we would like to support the **search for relations and actions**, and to support search via entailment.

Meaning-based Search

Traditional keyword-based protocols perform relatively well for queries that list entities or concepts, which are typically represented as nouns, since they often appear in the target documents. However, this approach performs poorly when the search is for **actions** or for **relations**, which are typically represented as verbs. This disparity is due to the variability in expressing actions and relations; identical meaning can be expressed in multiple ways.

In the context of search coverage, meaning-based search promises to revolutionize how we search for information. Using Relation Extraction and Textual Entailment, it will support semantic-based search and true **content-based access to information**. Such queries can only be satisfied when the search system plays an active role in reformulating the query and in semantically analyzing the retrieved candidate text.

For example, when searching for "*countries visited by the President of the United States over the past year*", the first page of documents returned by major search engines is unrelated, because the documents of interest may not contain the terms "*countries*", "*President of the United States*" or "*last year*" – even though the meaning of the query and expected results are clear.



Relation Extraction

Relation Extraction deals with **identifying interesting relationships between entities** in unstructured text. These relations are typically verb forms that describe actions by one entity that affect another. When a user queries for a relation, such as "what does Hyundai produce", he/she is looking for all documents that express, in some way, this relation, even though it can be expressed in many different ways, such as "Cars manufactured by Hyundai and Suzuki have the best P2P ratio".

Meaning-based Search involves both Relation Extraction and enabling search over the extracted relations. This poses unique challenges in finding paraphrases and entailments, indexing relations, and filtering based on similar/compatible entity types.

Textual Entailment

Textual entailment is the problem of determining if the meaning of a sentence or a short paragraph (Text) entails that of another (Hypothesis). This is a fundamental problem in natural language understanding that provides a broad framework for studying language variability and, aside from immediate applications in Questions Answering, Information Extraction and other Information Access tasks, is **ideally suited for searching over extracted relations.**

Publications

[1] M. Connor and D. Roth, *Context Sensitive Paraphrasing with a Single Unsupervised Classifier.* ECML, 2007.

[2] Quang Do, Dan Roth, Yuancheng Tu, *A Purely Lexical Approach to Textual Entailment*, ACL-HLT, 2008 (submitted)

Techniques

We have focused on **textual entailment using only "lightweight" lexical information**, developing similarity and relatedness metrics between concepts [2] and context sensitive metrics for relation descriptions [1]. Our textual entailment algorithm **performs better than standard keyword-based methods** on existing textual entailment and paraphrase-based corpora. [2]

Our work on **Relation Extraction** has focused on **reformulating the query** based on **context-based paraphrasing** [1] and on **modeling queries** using examples. Our initial evaluations show **promising results in both relation search and search using entailment**, with more relevant paraphrases being returned as search results as compared to simple keyword-based techniques.

Example

Query: "What countries did the U.S. President visit last year?"

Bad candidate: (top Google answer):
Ron Paul campaign site: "Campaign for President of the United States…"

Bad candidate: (top Yahoo! answer):
Democracy Now/Howard Zinn article: "…the US Ambassador to Australia, became partners with the future president…"

Good candidate: (wikipedia/George W. Bush):
"During a June 2007 visit to Albania, Bush was greeted with a 'rockstar reception'…"