

## New Search Paradigms to Facilitate Meaning based Information Retrieval

Quang Xuan Do, V.G.Vinod Vydiswaran, and Dan Roth  
University of Illinois at Urbana-Champaign, COE  
Multimodal Information Access and Synthesis Center  
(Dan Roth, Principle Investigator)

**Project Scope:** 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. For example, 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. 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*, as well as support *search via entailment*. Entailment-based search promises to revolutionize how we search for information; 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.

**Recent Progress:** We focused on textual entailment using only “lightweight” lexical information. Specifically, we developed similarity and relatedness metrics between concepts [2] and context sensitive metrics for relation descriptions [1]. We extended the metrics to include compound nouns and named entities. We modeled the relation search as a query-by-example, where the query is reformulated using context-based verb paraphrases. Our initial evaluations show promising results in both relation search and search using entailment, with much more relevant results as compared to simple keyword-based techniques.

**Future Plans:** We plan to further utilize all possible lexical features for entailment, primarily because they can be extracted with very light processing. Using a lightweight entailment system will help scale the *relation search* system to handle large text corpora. Further, we plan to explore techniques that will improve the indexing of entailment relations. Other plans include scaling-up similarity search over large data sets and modeling the user-assisted contextual relation search.

**Relevance to listed research areas:** This work applies novel techniques to garner information from large text data sets and is closely related to the *Advanced Data Analysis and Visualization* research area.

### Publications:

[1] M. Connor and D. Roth, Context Sensitive Paraphrasing with a Single Unsupervised Classifier. Proc. of the European Conference on Machine Learning (ECML) (2007)

[2] Quang Do, Dan Roth, Yuancheng Tu, *A Purely Lexical Approach to Textual Entailment*, The 46<sup>th</sup> Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, ACL-HLT, 2008 (submitted)