Information Systems 34 (2009) 753–765

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/infosys

Exploring models for semantic category verification

Dmitri Roussinov [a,*], Ozgur Turetken [b,1]

[a] Department of Computer and Information Sciences, University of Strathclyde, L13.29 Livingstone Tower, 16 Richmond Street, Glasgow G1 1XQ, United Kingdom
[b] Institute of Innovation and Technology Management, Ted Rogers School of Information Technology Management, Ryerson University, 575 Bay Street, Toronto, Ont., Canada M5G 2C5

ARTICLE INFO

Keywords:
Semantic category verification
Automated question answering
Text mining
World wide web
Search engines

ABSTRACT

Many artificial intelligence tasks, such as automated question answering, reasoning, or heterogeneous database integration, involve verification of a semantic category (e.g. "coffee" is a drink, "red" is a color, while "steak" is not a drink and "big" is not a color). In this research, we explore completely automated on-the-fly verification of a membership in any arbitrary category which has not been expected a priori. Our approach does not rely on any manually codified knowledge (such as WordNet or Wikipedia) but instead capitalizes on the diversity of topics and word usage on the World Wide Web, thus can be considered "knowledge-light" and complementary to the "knowledge-intensive" approaches. We have created a quantitative verification model and established (1) what specific variables are important and (2) what ranges and upper limits of accuracy are attainable. While our semantic verification algorithm is entirely self-contained (not involving any previously reported components that are beyond the scope of this paper), we have tested it empirically within our fact seeking engine on the well known TREC conference test questions. Due to our implementation of semantic verification, the answer accuracy has improved by up to 16% depending on the specific models and metrics used.

1. Introduction

Semantic verification is the task of automated verification of the membership in an arbitrary (not pre-anticipated) category, e.g. *red* is a *color*, *coffee* is a *drink*, but *red* is not a *drink*. While the problems arise in many domains, here we specifically explore its applications to online fact seeking, which is sometimes referred as open-corpus/open-domain question answering. Our approach builds on massive pattern matching which we believe models human linguistic practice of digesting evidence for categorical membership during a lifetime of learning process.

The goal of question answering is to locate, extract, and represent a specific answer to a user question expressed in natural language. Answers to many natural language questions (e.g. *What color is the sky*?) are expected to belong to a certain semantic category (e.g. *color* such as *blue, red, purple, etc.*), Those questions prove to be relatively difficult for current systems since the correct answer is not guaranteed to be found in an explicit form such as in the sentence *The color of the sky is blue*, but rather may need to be extracted from a sentence answering it implicitly, such as *I saw a vast blue sky above me*, in which a wrong answer "vast" has grammatically the same role as the correct answer "blue", and represents a property of the sky. However, *vast* refers to *size*, while we are looking for a *color*.

The currently popular approach to solving this "semantic" matching problem is through developing an extensive taxonomy of possible semantic categories [20]. This requires the anticipation of all possible questions, and hence substantial manual effort. Moreover, this

* Corresponding author. Tel.: +44 141 548 3706.
E-mail addresses: dmitri.roussinov@cis.strath.ac.uk (D. Roussinov), turetken@ryerson.ca (O. Turetken).
[1] Tel.: +141 6979 5000x2481.

0306-4379/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.is.2009.03.007

approach poses significant limitations, and although it works relatively well with more common categories (*cities, countries, organizations, writers, musicians*), it does not handle at least the following types of categories: (1) *Rare categories*: e.g. *What was the name of the first Russian astronaut to do a spacewalk? What American revolutionary general turned over West Point to the British?* (2) *Categories involving logic*: e.g. the question *What cities in Eastern Germany have been bombed during World War II?* involves a category defined as logical conjunction of being a *city* and being located in Eastern Germany. (3) *Vague categories*: e.g. the question *What industry is Rohm and Haas in?* involves a category *industry*, for which a simple Google search lists several definitions including "broad term for economic activity", a "sector", "people or companies engaged in a particular kind of commercial enterprise", and "businesses concerned with goods as opposed to services."

In this paper, we explore completely automated on-the-fly verification of a membership in a previously unanticipated category. Although, our algorithm can be used inside any other system, we have implemented and empirically evaluated it within our fact seeking engine, which has been available in a demo version online [42]. Our inspection of the 1000+ search sessions recorded by our demo reveals that approximately 20% of questions processed by the system have answers that are expected to belong to a specific semantic category, thus such systems can certainly benefit from semantic verification. The performance of our system was evaluated earlier [35,36] and found to be comparable with the other state-of-the-art systems (e.g. [14,22]) that are based on redundancy, rather than on extensive manually codified knowledge such as elaborate ontologies or rules for deep parsing. Contrary to the "knowledge-heavy" commercial systems, our system is entirely transparent: all the involved algorithms are described in prior publications, and thus, can be replicated by other researchers, which we believe makes this work superior to those reported on the "closed" (impossible to replicate) systems.

Through the work reported in this paper, we improve the semantic verification component of our system by moving beyond pure heuristics, and by building a model based on a logistic regression. Our hypotheses focus on (1) what variables contribute to the accuracy of answers, (2) what normalizing transformations are beneficial, and (3) if the improvements due to category verification are statistically and practically significant.

2. Literature review

The problems of automated verification of the membership in an arbitrary (not pre-anticipated) category exist in many domains including (1) *Automated Question Answering*: For example, the correct answer to the question *What soft drink has most caffeine?* should belong to the category "soft drink." (2) *Database federation*, where the automated integration of several heterogeneous databases requires matching an attribute in one database (e.g. having such values as *red, green,* and *purple*) to an attribute (e.g. *color*) in another database. (3) *Automated reasoning*, where the rules are propagated to all the subclasses of the superclass. (4) *Spellchecking* or *oddity detection* [17], where the substitution of a word with its hypernym (superclass) or hyponym (subclass) is considered legitimate while many other types of substitutions are not.

2.1. QA technology

The National Institute of Standards (NIST) has been organizing the annual Text Retrieval Conference (TREC) [39,40] since 1992, in which researchers and commercial companies compete in document retrieval and question answering tasks. The participating systems have to identify exact answers to so-called *factual* questions (or *factoids*) such as *who, when, where, what*, etc., list questions (*What companies manufacture rod hockey games?*), and definitions (*What is bulimia?*). In order to answer these questions, a typical participating system would: (a) transform the user query into a form it can use to search for relevant documents (web pages), (b) identify the relevant passages within the retrieved documents that may provide the answer to the question, and (c) identify the most promising candidate answers from the relevant passages. Most of the systems are designed based on techniques from natural language processing, information retrieval, and computational linguistics. For example, Falcon [20], one of the most successful systems, is based on a pre-built hierarchy of dozens of semantic types of expected answers (*person, place, profession, date, etc.*), complete syntactic parsing of all potential answer sources, and automated theorem proving to validate the answers.

In contrast to the natural language processing-based approaches, "shallow" approaches that use only simple pattern matching have recently been tried with good level of success. For example, the system from InsightSoft [38] won the 1st place in 2002 and the 2nd place in 2001 TREC competitions. The "knowledge-light" systems based on simple pattern matching and redundancy (repetitions of the answer on the Web), such as [14], also scored comparably.

Both NLP-based approaches and those that require elaborate manually created patterns have a strong advantage: they can be applied to smaller collections (e.g. corporate repositories) and still provide good performance. However, none of the known top performing systems has been made publicly open to the other researches for follow up investigations because of the expensive knowledge engineering required to build such systems and the related intellectual property issues. As result, it is still not known what components of these systems are crucial for their success, and how well their approaches would perform outside of the TREC test sets.

Meanwhile, the algorithms behind some of the systems that do not require extensive knowledge engineering, but still demonstrate reasonable performance, have been made freely available to public. Therefore replication of these systems and independent testing by other researchers is possible. We believe that from a research