Natural Language Processing Research

Speaker
Dr. Ashwin Ittoo, Asst-Prof, Management Information Systems, HEC-ULg

Contact: ashwin.ittoo@ulg.ac.be

Topics
1. Question-Answering Systems
2. Semantic Relation Extraction (if time permits)

Description

Question-Answering
In this talk, I will give an overview of my current research in Question-Answering (QA) systems. QA systems are considered as one of the major breakthroughs in Computer Science. In the recent years, they have received significant attention from academia and industry. For example, the IBM Watson QA system\(^1\) won the prize money of the Jeopardy show by outperforming human champions in answering quiz questions. Despite considerable progress, several challenges still exist in the design and implementation of QA systems. Most notably, the knowledge acquisition phase of classical QA systems relies on deep Natural Language Processing (NLP), such as syntactic parsing and co-reference resolution, to infer facts from huge volumes of texts. However, these sophisticated NLP techniques are not always accurate, leading to error propagation in the pipelined QA architecture, and ultimately compromising the overall performance. Furthermore, deep NLP is computationally expensive, which hampers the applications of QA systems in real-life, practical applications.

At the same time, the Linked Data (LD) initiative\(^2\) has led to the proliferation of a massive volume of structured, semantically-enriched data in ontologies published online by various providers. These ontologies span a wide spectrum of domains, encompassing medical, pharmaceutical, life sciences, general knowledge, music, movies, literature and science and technology. LD ontologies are collectively referred to as the largest knowledge base ever created in Artificial Intelligence. Thus, LD holds much promise for a variety of applications. However, the knowledge available from LD ontologies is still “dormant” in the sense that it is not easily accessible. The main reason is that the ontologies are encoded in specialized formats, such as RDF-S and OWL, and their querying requires expertise in formal logic and the sophisticated SPARQL query language.
In my research, I propose to develop a QA system that acquires its intelligence and reasoning capabilities directly from LD ontologies. The key contribution of such an endeavor is two-folds. First, it overcomes the challenges faced by extant QA systems in acquiring knowledge from text by alleviating the need for deep NLP. Second, it provides a natural language interface that enables users to pose free-formed questions to LD ontologies, overcoming the need for complex, formal SPARQL queries.

However, building a QA system over LD ontologies engenders several pertinent research questions that are yet to be fully addressed. The first question is how to resolve the lexical chasm between words/terms in the users’ questions and the linguistic realization of ontological concepts? Another question is how to resolve aliases, which are the same concepts but identified differently in different ontologies? Then, there is the issue of answering complex questions. These types of questions have been largely overlooked in extant QA research since determining their answers requires sophisticated reasoning/inferencing procedures that are non-trivial to perform over text data. In my presentation, I will elaborate on these challenges, and propose solutions, based on various linguistic theories and machine learning techniques, to address them.

**Semantic Relation Extraction**

If time permits, I will address another of my research area, which is Semantic Relation Extraction. Specifically, I will present a technique for learning causality relationships, between a causal agent and an effect, such as “hiv-aids”, “smoking-cancer” and “heated circuit board – blinking monitor”, from text. Compared to previous studies, the novelty of my technique is that it is minimally-supervised, alleviating the need for huge amounts of manually-annotated training data, which are not always readily available and are tedious to create. Furthermore, in addition to explicit causal relations, expressed by verbs such as “(to) cause” and its synonyms, my technique also discovers more implicit manifestations of causality, which are realized by expressions that do not have an explicit causal connotation, such as “destroy”, “create” or “rise”. Furthermore, the proposed technique achieves state-of-the-art performance in detecting causality relations across corpora of different genres, such as the open-domain Wikipedia collection and domain-specific corporate texts. This study has been published in Data and Knowledge Engineering (one of the top journals in Computer Science/Artificial Intelligence\(^3\)), and can be accessed at \texttt{http://dx.doi.org/10.1016/j.datak.2013.08.004}.

---

\(^1\) \url{http://www-03.ibm.com/innovation/us/watson/}
http://linkeddata.org/

Journal Ranking on Artificial Intelligence
http://www.scimagojr.com/journalrank.php?area=1700&category=1702&country=all&year=2012&order=sjr&min=0&min_type=cd