Some Challenges in the Design of Comparative and Evaluative Question Answering Systems

Nathalie Rose Lim
De La Salle University-Manila
2401 Taft Avenue, Philippines
Université Paul Sabatier
118 route de Narbonne, France
nats.lim@delasalle.ph

Patrick Saint-Dizier
IRIT-Université Paul Sabatier
118 route de Narbonne,
Toulouse, France
stdizier@irit.fr

Rachel Roxas
De La Salle University-Manila
2401 Taft Avenue,
Manila, Philippines
rachel.roxas@dlsu.edu.ph

Abstract

Comparative and evaluative question answering (QA) requires a detailed semantic analysis of comparative expressions and complex processing. Semantics of predicates from questions have to be translated to quantifiable criteria before extraction of information can be done. This paper presents some challenges faced in answering comparative and evaluative questions. An application on the domain of business intelligence is discussed.

1 Introduction

In the recently updated paper by Burger, et al. (2009), it is indicated that new types of questions like evaluative and comparative questions must be targeted in question answering (QA) systems. Evaluative refers to the consideration of at least one property or criteria over one or more entities and the computation of the associated values. Comparative refers to the evaluation of objects depending on one or more criteria and classifying those objects depending on the returned values. Included in comparative is the identification of the extreme, i.e., the superlatives, the topmost objects. In such cases, the focus of the questions is on the properties at stake in the evaluation, leading to the comparison. Thus, comparative and evaluative QA involves answering questions that require various forms of inference related to evaluation before an answer can be given. Since evaluation is necessary, the answer is not lifted from source text, as in the case of answering factoid, definition, or list questions. Instead, natural language answers will have to be constructed from the results of numeric and non-numeric evaluations of the criteria.

Currently, to our knowledge, there are no systems that answer comparative and evaluative questions. The closest applications to comparing or evaluating information are implemented through natural language database interfaces (Olawsky, 1989) and database queries (e.g., via SQL statements). In the former, the user is prompted to choose among a set of candidate interpretations of comparative expressions to indicate his intent. The comparisons are based on quantifiable predicates (i.e., those measurable by count, mass, or value). Using database queries restrict the possible questions that can be raised and is far less natural and user-friendly than using human language. It also does not allow producing cooperative responses.

Recent researches in linguistics on the semantics of comparatives and superlatives (Kennedy, 2006) can be used as a basis in answering comparative and evaluative questions. The next section discusses some challenges we have identified as crucial for the development of comparative and evaluative QA systems. We briefly propose some research directions we have explored or evaluated. We end this short document by a few illustrations from two applications we have worked on during the past year.

2 Challenges

The processes involved in classic components of a QA system are not only more complex but different for comparative and evaluative QA.

2.1 Question Analysis and Semantics of Comparatives

A question analyzer must identify the comparative expressions in the question and decompose it into meaningful constituents, among which are those properties that will be evaluated and the parameters of the comparison. Issues include:

- Identifying the type of comparison
Comparisons may be in relation to properties within the same object, degree of comparisons of the same property between different objects, or

Proceedings of the 2009 Workshop on Knowledge and Reasoning for Answering Questions, ACL-IJCNLP 2009, pages 15–18,
Suntec, Singapore, 6 August 2009.

different properties of different objects (Kennedy, 2006). In some simple situations, comparative relations in sentences can be extracted automatically via machine learning (Jindal and Liu, 2006). Their approach determines whether the expression is non-equal gradable, equative, or superlative. From this, the type of comparison may be determined from the semantics of the predicate and the properties of the objects through the pairability constraints. In our approach, we want to explore in more depth semantic and conceptual issues and their dependence to context, users, and domains.

- Determining semantic meaning and converting to quantifiable measures
The properties at stake in the comparison are embedded in the semantics of the words in the question, and possibly in the context that comes with the question. To date, there is obviously no widely available lexical resource containing an exhaustive list of comparative predicates, applied to precise terms, together with the properties involved. These can possibly be derived, to a limited extent, from existing resources like FrameNet or from an ontology where relationships between concepts and terms can be mapped. However, this is tractable for very simple situations, and in most cases, identifying those properties is a major challenge. We plan to explore, over restricted domains, ways to accurately identify those properties through different resources (like Generative Lexicon) and elaborate on inferential models to associate properties for evaluation.

- Determining limits, ranges, and values that are relative depending on the object
The standard of comparison (i.e., the value) associated to the predicate may be different based on the context, i.e., depending on the object that it is associated to and on the type of predicate. Properties of predicates may be underspecified and/or polysemic and would gain context only when associated with the object. One such predicate is *innovative*. The following are some properties that can be used to evaluate *innovative*.

- innovative product: type of product, number of entities interested in acquiring the product
- innovative company: strategy employed, type of product it produces
- innovative research: number of papers published on the same research, number of citations from other authors

To automatically determine the properties, including default values, to be used in the evaluation, other available sources indicating some range of values may be tapped, as is done in answer fusion (Girju, 2001). But rather than retrieving the partial answer, properties needed for evaluation must be retrieved or inferred. In terms of values, we have either numerical values (where comparisons are quite easy to handle) or textual values (that are often discrete). It is then necessary to define comparative scales along basic properties so that those values get ordered. This is a major challenge for our project.

- Processing superlatives and other forms of quantification related to comparisons
Superlatives and other forms of quantifications in connection with comparative expressions can also be used on top of the basic evaluative expressions. As the semantics of the predicate may encompass multiple properties, strict evaluation of these may trim the list prematurely. Consider the question:

- Which companies *take* the *most risk*?

Take most risk entails different dimensions from being *conservative*. In the context of business intelligence, evaluation could be in terms of the amount of investments, types of products invested in, the partners being taken, or all of these criteria. If a strict evaluation of all these criteria is done, the result may not be complete or accurate. We are exploring on relaxing the evaluation of multiple properties before determining the top results and on evaluating the superlative of each of the properties so as to identify which of the properties the object has not met.

2.2 Answer Determination

Only when the predicate/s is/are decomposed into properties can proper evaluation take place. We have two situations: either the QA system is connected to a database (which may have been constructed from natural language data as in the case of economic news) or it searches for the response on the Web. In the first case, the main challenge is to convert the concepts of the query into those of the conceptual schema of the database.

In the second case, relevant data must be searched on the Web. A straightforward procedure

consists of extracting keywords from the question, then getting results from search engines from which, via local grammars associated to properties, relevant values may be extracted. We already successfully conducted such an experiment for numerical data fusion (Moriceau, 2006).

2.3 Response Generation

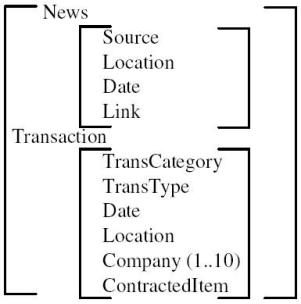
The answer cannot be lifted from the source text, thus a response generator component should be part of a comparative and evaluative QA system. As response is to be generated from the results of numeric and textual comparisons of the criteria, it is necessary to go through complex sentence generation, involving comparative expressions. In case the response is not direct, it is also necessary to elaborate adapted forms of cooperativity, by providing the user with adequate forms of explanations, elaborations, examples (of properties), and other relevant information. This is clearly a major challenge, since the quality of the response will reflect the overall credibility of the system.

3 Applications

We first carried out a relatively simple experiment on the business intelligence domain, where the criteria for evaluation are almost an exact science. The difficulty is to get the expertise in economics and to formulate it in terms of properties "visible" in the related economic news. An example question is given in (1).

1. `Which private biotech companies in Asia have the highest number of transactions from 2005 to 2008?`

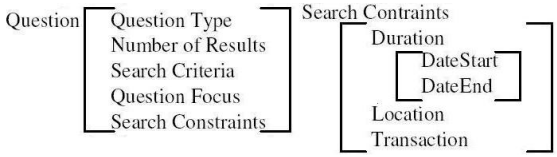
News articles are used as the source of information to answer these types of questions. They are factual, structured, and concise. They do not contain conflicting information, though there is the possibility of updates but the date is normally included in the information to provide temporal perspective. Rhetorical relations between sentences are being explored to give hints as to the relevance of information in the sentences. Semantic dependencies via thematic roles of arguments within each sentence are being considered to extract data. From the semantic dependency representation, a conceptual representation of these information is created using type-feature structure with the following information:



Location and Date are complex types containing info like country and month, respectively. TransCategory and TransType are transaction categories and its transaction subtype. There can be at most ten companies, where each contains information like the name and location of the company. The ContractedItem is also a complex type containing information like worth of the product.

To build this knowledge base, other web sources are used and a set of inferencing rules is developed to retrieve and store the required information.

Similarly, questions are represented semantically using thematic roles. Then, a conceptual representation is built to map the question focus with the answer from the type-feature representation of the news. To illustrate, the question type-feature structure contains the following:



For criteria or properties that are already in the conceptual representation, these are used in the evaluation and/or comparison. For question (1), occurrences of each company that fit the constraints (e.g., location Asia), are counted and the resulting values are compared to determine the top companies.

However, in (2), the sample question involves a non-directly translatable predicate.

2. `Does Company X take more risks than Company Y?`

Non-directly translatable predicates can be quantifiable by one criterion (e.g., active company: company with above-mean number of transactions), quantifiable by multiple criteria (e.g., company that take-risk: active company that has transactions every year, and has alliances every year but always with new partners or has unstable partners), polysemous (e.g., stable can mean ability to resist motion, steady in purpose, or established),

and/or underspecified (e.g., stable company vs. stable partner, though partner is also a company, the criteria is not the same. Stable company is an active company that may not have alliances every year or have alliances every year but always with old partners, whereas a stable partner is a company with alliances every year). There is also the issue of metonymy. In the context of company, the set of quantifiable properties associated to company could be number of employees, number of transactions, type of partners, and so on. Choosing which of these properties to associate to evaluate a predicate (like *stable*) is a challenge.

In this application, the categories, classifications, boundaries (what the term entails), and evaluation criteria of the terms are defined by an expert, so the result is consistent and objective. The challenge is to analyze the given information and convert it to machine tractable instructions. At present, set theory is used to define constraints and to generate the answer. It should be noted that it is one expert's interpretation of the terminologies used in the constraints. Others may have different criteria to associate with the predicates.

Other domains, like tourism, may be more challenging. Aside from information sources being not purely textual (i.e., some may be in tables or diagrams), the evaluation criteria for questions (3) and (4) may be subjective and may produce conflicting results. For example, value for money is subjective since certain amenities may not be important to the user. This can be resolved by prompting the user for additional criteria, by having a user profile, or by comparing with other entities (in this case, other hotels) to determine what is considered the norm (as a gauge to what is exceptional). It is also possible to generate different results based on the various criteria and present these to the user with explanations on the basis used.

3. `Which hotels in Singapore offer the most value for money for stay from August 28, 2009?`

4. `Which Asian cities are most kid-friendly?`

5. `Which hotels in Asia are most kid-friendly?`

As mentioned, the properties at stake in the evaluation could be different if the question focus was changed, as in the case of "kid-friendly" in question (5). In question (4), the criteria for a kid-friendly city could be one with avenues for fun and entertainment (like theme parks, zoos, parks) and a city with low crime rate (or specifically, low child abuse rate). On the other hand, a kid-friendly hotel would be one with amenities for supervised or planned activities, proximity to entertainment venues, larger rooms, or special menu for kids. The criteria or properties cannot be easily and reliably accessed from an ontology. Our challenge here is to elaborate means to get those properties. A direction we are investigating includes learning these properties from the web, but we may be faced with the recurrent problem of data sparseness, besides the fact that the web contains many erroneous statements.

Acknowledgments

The authors would like to thank Prof. Brigitte Gay of Ecole Superieure de Commerce - Toulouse for her invaluable inputs regarding comparative and evaluative questions in business intelligence.

References

John Burger, et al. 2009. *Issues, Tasks and Program Structures to Roadmap Research in Question & Answering.* Available in: www-nlpir.nist.gov/projects/duc/papers/qa.Roadmap-paper_v2.doc.

Roxana Girju. 2001. *Answer Fusion with Online Ontology Development.* In Students Research Workshop of the Second Meeting of the North American Chapter of the ACL. Available in: lyle.smu.edu/˜ roxana/papers/NAACL01.ps.

Nitin Jindal and Bing Liu. 2006. *Mining Comparative Sentences and Relations.* In Proceedings of the 21st AAAI Conference on Artificial Intelligence. AAAI Press, California,USA.

Christopher Kennedy. 2006. *Comparatives, Semantics Of.* In K. Allen (section editor) Lexical and Logical Semantics; Encyclopedia of Language and Linguistics, 2nd Edition. Elsevier, Oxford.

Véronique Moriceau. 2006. *Numeric Data Integration for Cooperative Question-Answering.* In Proceedings of the Knowledge and Reasoning for Language Processing Workshop (KRAQ 2006). ACL, Italy.

Duane Olawsky. 1989. *The Lexical Semantics of Comparative Expressions in a Multi-level Semantic Processor.* In Proceedings of the 27th Annual Meeting on ACL. ACL, USA.