PRECISE on ATIS: Semantic Tractability and Experimental Results

Ana-Maria Popescu **Alex Armanasu** **Oren Etzioni** **David Ko** **Alexander Yates**
University of Washington
Department of Computer Science and Engineering
Box 352350
Seattle, Washington 98195–2350
etzioni@cs.washington.edu

Abstract

The need for Natural Language Interfaces to databases (NLIs) has become increasingly acute as more and more people access information through their web browsers, PDAs, and cell phones. Yet NLIs are only usable if they map natural language questions to SQL queries *correctly* — people are unwilling to trade reliable and predictable user interfaces for intelligent but unreliable ones. We describe a reliable NLI, PRECISE, that incorporates a modern statistical paser and a semantic module. PRECISE provably handles a large class of natural language questions correctly. On the benchmark ATIS data set, PRECISE achieves 93.8% accuracy.

Introduction

Research on Natural Language Interfaces to databases (NLIs) has tapered off since the mid 1980's (Androutsopoulos, Ritchie, & Thanisch 1995). Yet more and more non-technical people access an increasingly complex array of databases through their web browsers and PDAs, making NLIs that reliably map English questions to SQL statements increasingly desirable.

PRECISE pushes the boundaries of NLI research in several ways. It is a "transportable" NLI in that it minimizes manual, database-specific configuration. Second, it incorporates the important advances made by statistical parsers over the last two decades, using semantic information to override syntactic information in cases where it knows how to correct parser mistakes. And third, PRECISE is built on a theoretical foundation that identifies a class of *Semantically Tractable* (ST) questions for which PRECISE is always correct, given correct syntactic and lexical information. Thus PRECISE is making progress, both theoretically and experimentally, on the goal of a reliable NLI.

Terminology and Theory

In an NLIDB, interpretations of a sentence are SQL statements. By giving a set of lexical constraints, semantic constraints, and syntactic constraints, we define what it means for an SQL statement to be a *valid interpretation* of a question. Our theoretical goal is to find classes of questions for which we can provably identify all the valid interpretations. While understanding natural language questions in general may be regarded as an "AI-complete" problem, questions like "What are the Chinese restaurants in Seattle?" are relatively easy to understand. Thus we have defined the class of ST questions, and we have shown that PRECISE always returns exactly the set of valid interpretations for any ST question, given correct syntactic and lexical information. (Etzioni *et al.* 2004)

PRECISE in Action

Consider the following question q_1 at the beginning of a dialog: "What are the flights from Boston to Chicago on Monday?"

Non-elliptical Sentence Interpretation. First, the system derives a set of word attachment constraints from q_1's parse tree. For example, "flight" is attached to "Boston" through preposition "from." The *tokenizer* produces a single tokenization of this question `(what, is, flight, from, boston, to, chicago,on,Monday)` and computes the token attachment function. PRECISE efficiently retrieves the set of matching database elements for every token from the *lexicon*. The *matcher* reduces the problem of satisfying the semantic constraints imposed by the valid interpretation definition to a graph matching problem and solves it efficiently with the Maxflow algorithm. The *query generator* takes the attachment information, the mappings of sentence tokens to database elements, the solution to the graph match, and a set of valid join paths between pairs of relations (as defined by the database administrator), and it constructs an SQL statement that obeys the semantic, syntactic, and lexical constraints identified by the inputs.

Elliptical Fragment Interpretation Now consider a follow-up question q_2, "from Boston to Atlanta?" PRECISE uses the *default*, *replacement* and *merging* operations to understand elliptical questions. The *default* operation consists of adding a wh-token to a tokenization of q_2, and it is the only one used when interpreting an elliptical sentence as a stand-alone utterance. The other two operations make use of the notion of *parallel* tokens: informally, tokens from two different sentences are parallel if they have the same token type and participate in *compatible* attachment constraints. The tokens `atlanta` and `boston` in q_2 and q_1, respectively, are not parallel because different

System Setup	PRECISE	PRECISEL
$Parser_{ORIG}$, *No Over-rides*	61.7%	60.1%
$Parser_{ORIG}$, *Over-rides*	89.5%	85.3%
$Parser_{ATIS}$, *No Over-rides*	92.2%	88.1%
$Parser_{ATIS}$, *Over-rides*	93.8%	89.1%
$Parser_{CORRECT}$, *Over-rides*	95.1%	91.3%

Table 1: **Impact of Parser Enhancements. The PRECISE column records the percentage of questions where the small set of SQL queries returned by PRECISE contains the correct query; PRECISEL refers to the questions correctly interpreted if PRECISE is forced to return exactly one SQL query.** $Parser_{ORIG}$ **is the original version of the parser,** $Parser_{ATIS}$ **is the version re-trained for the ATIS domain, and** $Parser_{CORRECT}$ **is the version whose output is corrected manually.** ***Over-rides*** **refer to the automatic use of semantic over-rides to correct parser errors.**

prepositions mediate their attachment, whereas `atlanta` and `chicago` are parallel because they have similar types and compatible attachment constraints. The *replacement* operation substitutes q_2 tokens for parallel q_1 tokens. In the merging operation, if there are tokens in q_2 that are not parallel to any in q_1 (or vice versa), they are kept as part of the final tokenization T'_{q_2}. Here are possible tokenizations T'_{q_2}, computed using the replacement and merging operations described above: T'_{q_2}=`what, fare, from, boston, to, atlanta, on, monday`, T'_{q_2}=`what, fare, from, boston, to, atlanta`.

Experimental Evaluation

Previously, we measured the practical impact of our theoretical framework by assessing the first version of our system on hundreds of natural language questions compiled by researchers at U.T. Austin (Tang & Mooney 2001) - we obtained 100% precision and approximately 80% recall. We then evaluated our system on the well-known ATIS dataset. The scoring set contains 445 context-independent sentences and 287 context-dependent sentences. There are 9 elliptical context-independent questions and 15 elliptical context-dependent questions. PRECISE was able to interpret ??? out of the 445 sentences in the A dataset. The rest were intractable either due to parser error (how many???) or due to intractability of ellipsis. 9 sentences were incorrectly interpreted. PRECISE correctly handles all 15 elliptical context-dependent sentences. The accuracy on transcribed text input on the A set is 93.8%, which is comparable to the response error rate of the best ATIS systems (see Table2).

Precise	AT&T	CMU
93.8%	96.2%	96.2%

Table 2: **Response error rate comparison: PRECISE vs. highest-performing ATIS systems.**

In another series of experiments, we tested how the behavior of PRECISE changed according to the quality of the syntactic information it was provided with (see Table 1). We measured PRECISE's accuracy using the Charniak parser, the Charniak parser retrained on questions from the ATIS domain, and manually corrected syntactic information. Overall, the results improved from 61.7% accuracy to 95.1% accuracy as the parser improved.

Related Work

There have been a great number of systems built for the ATIS domain. Semantic parser-based systems (Seneff 1992; Ward & Issar 1996) used manually-crafted, context-free semantic rules to fill slots in semantic frames. Stoachastic model-based systems (Levin & Pieraccini 1995) require a fully annotated corpus to build a reliable model. PRECISE's, on the other hand, uses a set of domain-independent semantic constraints and existing parsing technology so that it is more portable than both kinds of systems while performing comparably on ATIS.

Mooney's learning approach to database interfaces requires a great deal of hand tagged sentences, and it emphasizes recall over precision by attempting to interpret sentences with unknown words. As far as handling ellipsis is concerned, most previous work (such as (Crouch 1995; Dalrymple, Shieber, & Pereira 1991)) deals with verb phrase ellipsis and only tangentially mentions noun phrase elliptical fragments. (Carbonell 1983) requires a semantic grammar, so the application cannot easily be ported to another domain and processes only a subset of the elliptical fragments PRECISE can handle.

References

Androutsopoulos, I.; Ritchie, G. D.; and Thanisch, P. 1995. Natural Language Interfaces to Databases - An Introduction. In *Natural Language Engineering, vol 1, part 1*, 29–81.

Carbonell, J. 1983. Discourse pragmatics and ellipsis resolution in task-oriented natural language interfaces. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, 164–168.

Crouch, R. 1995. Ellipsis and quantifi cation: A substitutional approach. In *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics*, 229–236.

Dalrymple, M.; Shieber, S.; and Pereira, F. C. N. 1991. Ellipsis and higher-order unifi cation. *Linguistics and Philosophy* 14(4):399–452.

Etzioni, O.; Popescu, A.; Armanasu, A.; Yates, A.; and Ko, D. 2004. Semantically tractable questions and follow-up questions in the atis data. In *Submitted to Coling-2004*.

Levin, E., and Pieraccini, R. 1995. Chronus, the next generation. In *Proceedings of the DARPA Speech and Natural Language Workshop*, 269–271.

Seneff, S. 1992. Robust parsing for spoken language systems. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.

Tang, L., and Mooney, R. 2001. Using multiple clause constructors in inductive logic programming for semantic parsing. In *Proceedings of the 12th European Conference on Machine Learning (ECML-2001), Freiburg, Germany*, 466–477.

Ward, W., and Issar, S. 1996. Recent improvements in the cmu spoken language understanding system. In *Proceedings of the ARPA Human Language Technology Workshop*, 213–216.