Natural Language Data Management and Interfaces: Recent Development and Open Challenges

Yunyao Li
IBM Research - Almaden
yunyaoli@us.ibm.com

Davood Rafiei
Department of Computing Science
University of Alberta
drafiei@ualberta.ca *

ABSTRACT

The volume of natural language text data has been rapidly increasing over the past two decades, due to factors such as the growth of the Web, the low cost associated to publishing and the progress on the digitization of printed texts. This growth combined with the proliferation of natural language systems for search and retrieving information provides tremendous opportunities for studying some of the areas where database systems and natural language processing systems overlap. This tutorial explores two more relevant areas of overlap to the database community: (1) managing natural language text data in a relational database, and (2) developing natural language interfaces to databases. The tutorial presents state-of-the-art methods, related systems, research opportunities and challenges covering both areas.

1. MOTIVATION AND OVERVIEW

"If we are to satisfy the needs of casual users of data bases, we must break through the barriers that presently prevent these users from freely employing their native languages" [17]. Codd made those comments when relational databases were just taking off and machine resources were too limited to process natural language queries and text. Today natural languages play a much bigger role in our daily interactions with machines and we have a larger set of resources at our disposal, in terms of the processing power and public data sets and knowledge bases (e.g. Wikipedia, Yago), to build and train our models; furthermore there is a growing number of tools that can help with processing text, ranging from part-of-speech taggers, syntactic parsers, semantic role labeling tools, etc. The success of Watson [23] at Jeopardy has further ignited the interest in natural language text. This development has two implications as far as database research is concerned. First, we are amassing natural language text in sizes that we have not seen before and the sheer volume of information encoded in text and its relationships to data in our relational databases is too great to be ignored. Second, there is a huge opportunity to push database systems more in the direction of realizing Codd's vision of "rendezvous". This tutorial reviews the state-of-the-art methods, recent progress, research opportunities and challenges in two interrelated and timely topics of building natural language interfaces to databases and managing natural language text. The intended length of the tutorial is 3 hours.

2. MANAGING NATURAL LANGUAGE TEXT DATA

Natural language text may not be the best medium for expressing facts and describing relationships but the truth of the fact is that much of human knowledge and everyday information is written and communicated in natural language text. Traditionally this data has been stored outside databases and processed using natural language processing tools, in isolation from other data sources. There are many scenarios where natural language text co-exists and is queried with more structured data. Here are some scenarios.

- We have a collection of medical articles and want to find treatments for a disease and their success rates as reported in those articles.
- We are given the health records of a set of patients who have gone under a surgery and want to find post-operative complications as reported in the patient records.
- We have a few candidates running in a federal election and want to gauge the degree of support they are getting in social media (e.g. Twitter) and the contexts in which their names are mentioned.
- We are reporting products that satisfy some user needs and want to include some statistics or analysis of the sentiments on each product from its reviews.

A unifying theme in all these scenarios is that (1) text sources are queried and analyzed in granularities smaller than a document, and (2) text sources are queried in conjunction with more structured data which may be available either as meta-data of the same text sources (e.g. the poster of a tweet, the date it is posted, the user who retweeted it, etc.) or from different sources (e.g. the names of candidates and their parties). There is a large range of applications with the same or similar data requirements that can

*Yunyao Li and Davood Rafiei contributed equally to this work.

SIGMOD'17, May 14-19, 2017, Chicago, IL, USA
 ISBN 978-1-4503-4197-4/17/05…$15.00
DOI: http://dx.doi.org/10.1145/3035918.3054783

benefit from a possible integrated solution. In general, IR approaches are less useful when text is queried in small granularities such as a sentence or joined with structured data, hence a viable choice is to store natural language text in a relational database. The latter choice offers some benefits in terms of opportunities for query optimization and efficient query processing as well as allowing easier development of more complex applications. However there are two major challenges that hinder this development: (1) querying natural language text, (2) transforming natural language text to a format that can be easily queried, aggregated and joined with other sources. Other challenges that arise in both cases include variations in entity naming and referencing and differences due to synonyms and paraphrasing, text formatting, misspelling, etc. Despite these challenges, there have been major progress in both areas.We will review the progress made in those areas and also discuss some of the challenges.

2.1 Querying natural language text

Natural language text has a grammar, and the grammar of a sentence can be treated as a schema with the facts and relationships expressed in the sentence as instances, very much similar to schemas and instances in a relational database. This view of text allows many database concepts to be applied to documents in natural language text. We will review the relevant literature on text-pattern and tree-pattern queries [40, 39, 16, 11, 48] and some of the index structures that are applicable [7, 15, 6]. Two challenges here are (1) the presence of rewriting or paraphrasing relationships between text patterns [34, 49], and (2) the uncertainty associated with schema and data values [8]. We will review the progress made and some of the challenges.

If we treat natural language text as simply text, general approaches for querying text databases can be applied. These approaches range from simple keyword queries to approaches that use the document structure. A related line of research is the work inspired by the Oxford English Dictionary (OED) which resulted in some data models and languages for querying the grammar-based structure of documents [24, 43]. Recent related work includes the approaches on integrating text and relational data, which can be divided into *tight integration* and *loose integration*. Chu et al. [14] present a tight integration approach that incrementally queries the structure in text, and as more queries are processed, more structure is extracted, allowing a richer set of queries. There has been more studies on a *loose integration* of text and relational data where text sources are managed by a text search engine and are joined with relational data at the query time. Under a loose integration scheme, different probing strategies are studied and various cost models and query processing and optimization techniques are developed [12, 2, 26, 1]. A typical workload is entity extraction where a set of entities is stored in a relational database, and the goal is to efficiently retrieve the mentions of those entities in a set of documents [2]. This line of work on text can be directly applied to natural language text; however, natural language text has a structure (as discussed in this and next subsections) and offers more opportunities for querying and query processing. We will review this line of work in the context of some of those opportunities and challenges.

2.2 Transforming natural language text

Natural language text may be transformed into a formal meaning representation language with more powerful querying capabilities. The development in this area, even though not that old, has been very promising. In a pioneering work, Moldovan et al. [36, 20] map natural language text to logic forms and axioms. Based on this mapping, the authors develop tools which are successfully employed in a Question Answering (QA) system deployed by Language Computer Corporation (LDC) [1]. With recent development around knowledge graphs and RDF, there has been more progress in translating natural language utterances to predicates in Freebase [28, 10], Yago2 [50] and RDF [18]. These translation tools, better known as semantic parsers, use a combination of techniques (such as manual rules [50], distant supervision [28], schema matching [10], datasets such as Web text and knowledge base [5], etc.) and have been deployed in different domains such as mapping user queries in English or bag of words to knowledge base queries [41] and translation of open domain database queries [25].

The database community has been contributing to this line of research in the areas of querying (e.g. [37]), improving (e.g. [13]), expanding (e.g. [38]) and accelerating the construction of (e.g. [45]) knowledge bases that may be constructed from natural language text sources. For example, Chen and Wang [13] present a probabilistic approach to infer the missing values, and Nakashole et al. [38] report a system that can harvest high quality knowledge from natural language text sources. We will review the recent progress in this area, some of the data management challenges and contributions, and the research opportunities we see the database community can contribute.

3. DEVELOPING NATURAL LANGUAGE INTERFACES TO DATABASES

Natural language interfaces have been regarded as the holy grail for query interface to databases. An ideal natural language interface to databases (NLIDB) enables users to pose arbitrarily complex ad-hoc queries against databases and obtain precise information back with minimal effort. It requires no knowledge of any formal query language, database schema, or the exact terminology of the underlying data. Not surprisingly, the emerging democratization of data makes NLIDBs even more appealing than before. However, despite years of research efforts, NLIDB largely remains an open research question. Two major open challenges hinder the wide adaption of NLIDBs: (1) natural language understanding; and (2) query translation. We survey the advancements in addressing these research challenges and discuss notable examples developed since 2000, such as PRECISE [35], GeoDialogue [9], NaLIX [33, 32]/DaNaLIX [22], NaLIR [30, 31], NLPQC [46], NL_2CM [3], and ATHANA [42], including some discussed in [21].

3.1 Natural language understanding

Natural language understanding refers to the capability of parsing a natural language query, usually in the form of one single natural language sentence, into a data structure that represents the syntactical and/or semantic structure of the query. Natural language understanding is the foundation of any NLIDB system. Unfortunately, generic natural

[1] http://www.languagecomputer.com

language understanding remains an open research question by itself. Moreover, parsers used by NLIDBs are usually trained with open-domain news corpus, while the typical queries in NLIDBs are domain-specific questions. As a result, parsers tend to make mistakes when parsing natural language queries and cause issues for the subsequent operations in NLIDBs. We identify two dimensions that distinguish different approaches towards NLIDB, in terms of natural language understanding: *scope of natural language support* and *parser error handling*.

3.1.1 Scope of natural language support

The scope of a NLIDB may be characterized in terms of the types of natural language queries it supports.

Ad-hoc Natural Language Queries An ideal NLIDB system should support ad-hoc natural language queries. Not surprisingly, some NLIDBs (such as NaLIR [30], [4]) aim to support ad-hoc natural language queries. Unfortunately, parsing ad-hoc natural language queries remains an open problem. As such, NLIDBs supporting ad-hoc natural language queries have to heavily rely on parser error handling.

Controlled Natural Language Queries In practice, NLIDB systems often limit their scope of natural language support to reduce potential parser errors. For instance, PRECISE [35] defines the notion of semantic tractability and identifies a subset of natural language queries that can be precisely translated into SQL. NaLIX [33, 32] limits natural language queries to a controlled subset based on a predefined grammar. Similarly, NLPQC [46] accepts queries based on a domain-specific template. We will discuss how such systems overcome the two common major challenges of (1) ensuring the expressiveness of the controlled language, and (2) helping users to understand and learn and to effectively use the controlled language.

The scope of a NLIDB may also be characterized based on whether it is stateless or stateful (i.e. conversational).

Conversational An ideal NLIDB should be conversational. In another word, it should allow users to ask follow-up questions based on previous queries and provide appropriate answers based on the context. Notable example of conversational NLIDB include GeoDialogue [9] and NaLIX [33], and more recently [19].

Stateless Most existing NLIDBs are stateless. In a stateless NLIDB, queries are handled independently from each other. However, even in a stateless NLIDB, prior queries may still be used to improve the NLIDB (e.g. improve parser error handling). In addition, the style of user interaction of a stateless NLIDB may also appear to be conversational (e.g.ANNESAH [44]).

3.1.2 Parser error handling

Most work on NLIDB [27, 3] directly leverage existing work on natural language understanding and ignore potential parsing errors with a few notable exceptions.

Auto-Correction One approach is to automatically detect and correct parse errors before query translation. We will discuss how some of the NLIDBs (e.g. PRECISE [35] and DaNaLIX [22]) remedy this issue by detecting certain types of parser errors (e.g. based on external semantic information) and correct them automatically.

Interactive Correction Another common approach is to leverage user interaction and correct parser errors in an interactive fashion. We will discuss the two common approaches towards interactive correction: (1) Query reformulation, where users are asked to reissue the current query into one that can be correctly handled by the parser (e.g. NaLIX [33, 32]); and (2) Parse tree correction, where users are asked to correct parser errors so that the current query can be correctly understood by the system (e.g. NaLIR [30]).

3.2 Query translation

Query translation is the process of translating a parsed natural language query into a correct formal query against the underlying database. It is part of most NLIDBs, and the major challenges here are *bridging the gap between the parsed queries and the underlying data* and *generating formal queries*.

3.2.1 Bridging the Semantic Gap

One important challenge facing NILDBs is to bridge the gap between the user queries and the underlying data. Furthermore, users of a NLIDB typically have no precise knowledge of the underlying data. As a result, there often exists a mismatch between a user query and the underlying data. We will review different approaches towards bridging the gap, from simple synonym expansion [29] to carefully designed conversational UIs (e.g. [44]).

3.2.2 Query Construction

The construction of queries in a formal language from natural language queries can be done via machine learning [47, 4] or by constructing formal queries from parsed queries, potentially after resolving ambiguity and augmenting with additional information (e.g. [3, 42]). We will examines both approaches in our tutorial.

3.3 Relationships to question answering

Both NLIDBs and Question Answer (QA) systems take as input a question formulated in natural language and must interpret the question in order to answer it correctly. There are key difference between them: (1) the underlying data of NLIDBs is structured or semi-structured, while that of QA systems is unstructured; and (2) the queries supported by NLIDBs are typically much more expressive than those supported by QA systems. We will discuss the similarities and differences between them and outline challenges to synergy the two.

4. OPEN CHALLENGES AND OPPORTUNITIES

Further to some of the challenges and opportunities listed throughout this proposal, there are two major open research challenges on natural language data management and interfaces that are relevant to the database community.

Mobile Natural Language Data Management With the ubiquity of smart mobile devices, the scale and complexity of natural language data is growing fast, while the needs for conversational, situation-aware NLIDBs has never been stronger; this brings new challenges in natural language data management.

Unified Data Management Data in the real world are often mixtures of structured, semi-structured and unstructured data. To manage such data effectively, one major challenge is to build data management systems supporting heterogeneous data models, ranging from less-structured

data such as natural languages text to more-structured relational data. Another interesting opportunity is to build a NL interface for such a data management system by bringing techniques for NLIDBs and QA systems together and leverage information from both structured data and unstructured text to provide answers that are not possible before using one single type of data alone. Watson [23] is a successful example for such a system, but building such a system for arbitrary domains remains an open question.

To summarize, this tutorial overviews a multitude of problems in natural language data management and interfaces to DB, and discusses relevant, state-of-the-art techniques. It is our hope that this tutorial highlights some of the challenges ahead while helping researchers, data management developers and architects, as well as corporate stakeholders gain insight and better contribute to the field.

5. INTENDED AUDIENCE

This tutorial is intended for a wide scope of audience, including both database researchers and developers working on various topics from knowledge base creation and querying, data integration, query interfaces, and database applications. By learning the challenges, current solutions and future directions related to managing natural language data management and interfaces, these researchers and developers can better utilize their expertise and contribute to building better systems to address these challenges. This tutorial is also intended to benefit researchers and developers who are new to the topic, and help them quickly gain a comprehensive picture of the field.

6. ABOUT THE PRESENTERS

Yunyao Li is a research manager and research staff member at IBM Research - Almaden. She is also an IBM Master Inventor and a member of IBM Academy of Technology. She is interested in designing, developing, and analyzing large-scale systems that are usable by a wide spectrum of users. Her current research focuses on scalable natural language processing and knowledge engineering. She received her Ph.D. from the University of Michigan, Ann Arbor.

Davood Rafiei is an associate professor of Computer Science and a member of database research group at the University of Alberta. He obtained his M.Sc. from the University of Waterloo and his Ph.D. from the University of Toronto. His areas of interest include integrating natural language text with relational data and Web information retrieval. Davood has been a visiting scientist at Google, Kyoto University and the University of Paris V.

Acknowledgments

The authors wish to thank Dekang Lin and Lucian Popa for their comments on an earlier draft of this manuscript. Davood Rafiei's research is supported by the Natural Sciences and Engineering Research Council of Canada.

7. REFERENCES

[1] Eugene Agichtein and Luis Gravano. Querying text databases for efficient information extraction. In *Proc. of the ICDE Conference*, pages 113–124, Bangalore, India, March 2003.

[2] Sanjay Agrawal, Kaushik Chakrabarti, Surajit Chaudhuri, and Venkatesh Ganti. Scalable ad-hoc entity extraction from text collections. *PVLDB*, 1(1):945–957, 2008.

[3] Yael Amsterdamer, Anna Kukliansky, and Tova Milo. A natural language interface for querying general and individual knowledge. *PVLDB*, 8(12):1430–1441, August 2015.

[4] H Bais, M Machkour, and L Koutti. Querying database using a universal natural language interface based on machine learning. In *IT4OD*, 2016.

[5] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *Proc. of the EMNLP Conference*, volume 2, page 6, 2013.

[6] Elisa Bertino, Beng Chin Ooi, Ron Sacks-Davis, Kian-Lee Tan, Justin Zobel, Boris Shidlovsky, and Daniele Andronico. *Indexing techniques for advanced database systems*, volume 8. Springer Science & Business Media, 2012.

[7] Michael J Cafarella and Oren Etzioni. A search engine for natural language applications. In *Proc. of the WWW conference*, pages 442–452. ACM, 2005.

[8] Michael J. Cafarella, Christopher Re, Dan Suciu, and Oren Etzioni. Structured querying of web text data: A technical challenge. In *Proc. of the CIDR Conference*, pages 225–234, Asilomar, CA, January 2007.

[9] Guoray Cai, Hongmei Wang, Alan M. MacEachren, and Sven Fuhrmann. Natural conversational interfaces to geospatial databases. *Transactions in GIS*, 9(2):199–221, 2005.

[10] Qingqing Cai and Alexander Yates. Large-scale semantic parsing via schema matching and lexicon extension. In *ACL*, pages 423–433. Citeseer, 2013.

[11] Angel X Chang and Christopher D Manning. Tokensregex: Defining cascaded regular expressions over tokens. Technical Report CSTR-2014-02, Department of Computer Science, Stanford University.

[12] Surajit Chaudhuri, Umeshwar Dayal, and Tak W Yan. Join queries with external text sources: Execution and optimization techniques. In *ACM SIGMOD Record*, pages 410–422, San Jose, California, May 1995.

[13] Yang Chen and Daisy Zhe Wang. Knowledge expansion over probabilistic knowledge bases. In *Proc. of the SIGMOD conference*, pages 649–660. ACM, 2014.

[14] Eric Chu, Akanksha Baid, Ting Chen, AnHai Doan, and Jeffrey Naughton. A relational approach to incrementally extracting and querying structure in unstructured data. In *Proc. of the VLDB Conference*, 2007.

[15] P. Chubak and D. Rafiei. Index Structures for Efficiently Searching Natural Language Text. In *Proc. of the CIKM Conference*, 2010.

[16] Pirooz Chubak and Davood Rafiei. Efficient indexing and querying over syntactically annotated trees. *PVLDB*, 5(11):1316–1327, 2012.

[17] E.F. Codd. Seven steps to rendezvous with the casual user. In *IFIP Working Conference Data Base Management*, pages 179–200, 1974.

[18] Francesco Draicchio1 and Aldo Gangemi. Fred: From natural language text to rdf and owl in one click. In *Extended Semantic Web Conference*, pages 263–267, 2013.

[19] Eduardo M. Eisman, María Navarro, and Juan Luis Castro. A multi-agent conversational system with heterogeneous data sources access. *Expert Syst. Appl.*, 53:172–191, 2016.

[20] Dan Moldovan et al. LCC tools for question answering. In *TREC*, 2002.

[21] Rodolfo A. Pazos R. et al. Natural language interfaces to databases: An analysis of the state of the art. *Recent Advances on Hybrid Intelligent Systems*, 451:463–480, 2013.

[22] Yunyao Li et al. Enabling domain-awareness for a generic natural language interface. In *AAAI*, pages 833–838, 2007.

[23] David A. Ferrucci. Introduction to "this is watson". *IBM Journal of Research and Development*, 56(3):1, 2012.

[24] Gaston H. Gonnet and Frank Wm. Tompa. Mind your grammar: a new approach to modelling text. In *Proc. of the VLDB Conference*, pages 339–346, Brighton, England, September 1987.

[25] Carolin Haas and Stefan Riezler. Responsebased learning for machine translation of opendomain database queries. In *Proc. of NAACL HLT*, pages 1339–1344, 2015.

[26] Alpa Jain, AnHai Doan, and Luis Gravano. Optimizing SQL queries over text databases. In *Proc. of the ICDE Conference*, pages 636–645, Cancun, Mexico, April 2008.

[27] Rohini Kokare and Kirti Wanjale. A natural language query builder interface for structured databases using dependency parsing. *International Journal of Mathematical Sciences and Computing*, 1(4):11–20, November 2015.

[28] Jayant Krishnamurthy and Tom M Mitchell. Weakly supervised training of semantic parsers. In *Proc. of the EMNLP Conference*, pages 754–765. Association for Computational Linguistics, 2012.

[29] Nicolas Kuchmann-Beauger and Marie-Aude Aufaure. A natural language interface for data warehouse question answering. In *Natural Language Processing and Information Systems*, volume 6716, pages 201–208. 2011.

[30] Fei Li and H. V. Jagadish. Constructing an interactive natural language interface for relational databases. *PVLDB*, 8(1):73–84, 2014.

[31] Fei Li and H. V. Jagadish. Understanding natural language queries over relational databases. *SIGMOD Record*, 45(1):6–13, June 2016.

[32] Yunyao Li, Huahai Yang, and H. V. Jagadish. Constructing a generic natural language interface for an XML database. In *Proc. of the EDBT Conference*, pages 737–754, 2006.

[33] Yunyao Li, Huahai Yang, and H. V. Jagadish. Nalix: A generic natural language search environment for XML data. *ACM Trans. Database Systems*, 32(4), 2007.

[34] Dekang Lin and Patrick Pantel. Dirt - discovery of inference rules from text. In *Proc. of the KDD Conference*, pages 323–328, 2001.

[35] Ana maria Popescu et al. Modern natural language interfaces to databases: Composing statistical parsing with semantic tractability. In *Proc. of the COLING Conference*, 2004.

[36] Dan Moldovan and Vasile Rus. Logic form transformation of wordnet and its applicability to question answering. In *Proc. of the ACL Conference*, pages 402–409, 2001.

[37] Davide Mottin, Matteo Lissandrini, Yannis Velegrakis, and Themis Palpanas. Exemplar queries: Give me an example of what you need. *Proc. of the VLDB Endowment*, 7(5):365–376, 2014.

[38] Ndapandula Nakashole, Martin Theobald, and Gerhard Weikum. Scalable knowledge harvesting with high precision and high recall. In *Proc. of the WSDM Conference*, pages 227–236. ACM, 2011.

[39] Davood Rafiei and Haobin Li. Data extraction from the web using wild card queries. In *Proc. of the CIKM Conference*, pages 1939–1942, 2009.

[40] Deepak Ravichandran and Eduard Hovy. Learning surface text patterns for a question answering system. In *Proc. of the ACL Conference*, 2002.

[41] Siva Reddy, Oscar Täckström, Michael Collins, Tom Kwiatkowski, Dipanjan Das, Mark Steedman, and Mirella Lapata. Transforming dependency structures to logical forms for semantic parsing. *Transactions of the Association for Computational Linguistics*, 4:127–140, 2016.

[42] Diptikalyan Saha, Avrilia Floratou, Karthik Sankaranarayanan, Umar Farooq Minhas, Ashish R. Mittal, and Fatma Özcan. Athena: An ontology-driven system for natural language querying over relational data stores. *PVLDB*, 9(12):1209–1220, August 2016.

[43] Airi Salminen and Frank Tompa. PAT expressions: an algebra for text search. *Acta Linguistica Hungarica*, 41(1):277–306, 1994.

[44] K Shabaz, Jim D O'Shea, Keeley A Crockett, and A Latham. Aneesah: A conversational natural language interface to databases. In *World Congress on Engineering*, pages 227–232, 2015.

[45] Jaeho Shin, Sen Wu, Feiran Wang, Christopher De Sa, Ce Zhang, and Christopher Ré. Incremental knowledge base construction using deepdive. *Proc. of the VLDB Endowment*, 8(11):1310–1321, 2015.

[46] Niculae Stratica, Leila Kosseim, and Bipin C. Desai. Using semantic templates for a natural language interface to the cindi virtual library. *Data and Knowledge Engineering*, 55(1):4–19, October 2005.

[47] Lappoon R Tang and Raymond J Mooney. Using multiple clause constructors in inductive logic programming for semantic parsing. In *European Conference on Machine Learning*, pages 466–477, 2001.

[48] Marco A Valenzuela-Escarcega, Gustave Hahn-Powell, and Mihai Surdeanu. Odin's runes: A rule language for information extraction. In *Proc. of the Language Resources and Evaluation Conference (LREC)*, 2016.

[49] Wei Xu. *Data-driven approaches for paraphrasing across language variations*. PhD thesis, New York University, 2014.

[50] Mohamed Yahya, Klaus Berberich, Shady Elbassuoni, Maya Ramanath, Volker Tresp, and Gerhard Weikum. Natural language questions for the web of data. In *Proc. of the EMNLP Conference*, pages 379–390. Association for Computational Linguistics, 2012.