Multilingual Database Access in the Query Converter

Aarne Ranta\textsuperscript{1,2}

\textsuperscript{1} Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden  
\texttt{aarne@chalmers.se}

\textsuperscript{2} Department of Computer Science and Engineering, University of Gothenburg, Gothenburg, Sweden

Abstract. This paper is a report on work in progress on a piece of software that provides natural language access to databases. The functionalities include translating between texts and entity-relationship models as well as between natural language queries and SQL. The long-term goal of the work is to develop an accurate, generalizable, and scalable technique for natural language access to structured data.

Keywords: Controlled language · Entity-relationship models · Grammatical framework · Query language

1 The Mission

Natural language access to digital data is an old idea, dating back at least to 1959\textsuperscript{1}. However, in spite of all efforts, a general, scalable, and intuitive query system in natural language still doesn’t exist. The alternatives available for most uses are SQL-like formal notations, hierarchic menus, and string-based search.\textsuperscript{2}

But the topic of natural language question answering is active again, with systems like Apple’s Siri and Google’s Voice Search promising to change our practices of information search. Wolfram Alpha\textsuperscript{3} is a slightly less general system, but it gives access to sophisticated structured queries in those domains that it covers. The difficulty lies in combining precision with coverage: how to extend accurate query systems beyond limited domains?

In this paper, we will introduce yet another initiative in natural language based data management. It is a CNL (Controlled Natural Language) approach implemented in GF (Grammatical Framework)\textsuperscript{4}. Our goal is to develop a scalable, generalizable, and adaptable approach to accessing structured data. This includes

---


\textsuperscript{2} SQL was actually meant to be usable by non-programmers and resembles natural language with its COBOL-like syntax.

\textsuperscript{3} https://www.wolframalpha.com/.

\textsuperscript{4} http://grammaticalframework.org.
– translating between formal and natural language queries,
– adapting the translation to new domains and their concepts with a minimal effort,
– porting translation systems to new languages via lexical annotations.

The work has emerged as a part of qconv, Query Converter, an open-source project developing teaching software for an undergraduate database course. The software follows closely the first half of a standard text book. Thus qconv supports exercises and experiments related to textbook concepts: E-R (Entity-Relationship) modelling (graphics and text), functional dependencies and normalization, SQL parsing and interpretation, relational algebra parsing and interpretation, and the XML data format. The idea of qconv is to link all these components together, so that, in particular, natural language access via domain-adapted concepts works across different representations.

The natural language modules of qconv build on earlier work on queries in GF, in particular, on a CNL called YAQL used for natural language SPARQL queries. Working on SQL and the relational data model presents some new challenges, because of its rich hierarchic structure and its varieties of usage and idioms. What we show here is work in progress, yet promising enough to justify a first report and a demo in a workshop.

2 First Results

CNL is currently used in qconv in two places: E-R modelling and SQL queries. Here is an example E-R model with the corresponding text:

The CNL is inspired by the mapping between E-R objects and natural language categories proposed by Peter Chen, the inventor of E-R modelling. In qconv,

---

Chen’s mappings have been slightly generalized so that for instance relationships don’t need to be transitive verbs but can be relational expressions more generally. The pedagogical use of the CNL is to serve as an intermediate between an often messy informal text and the formal model. The students are advised to paraphrase the text in the CNL, which makes it easier to draw the model. One could also use the translation in the inverse direction: generate the text from a given model to understand the model better.  

The query CNL in qconv has a base grammar, which matches the logical structure of queries compositionally. This base grammar is extended with NLG-style (Natural Language Generation) optimizations. Domain-specific terminology for tables and attributes can be inherited from E-R models but also added separately. In particular, it is possible to add domain-specific idioms as a part of a domain lexicon. Here are some example queries and their English and Swedish translations (from a domain that extends the above E-R model).

```
SELECT capital FROM countries WHERE name='Sweden'
what is the capital of Sweden
vad är Sveriges huvudstad

SELECT name FROM Countries WHERE currency='EUR'
which countries have EUR as currency
vilka länder har EUR som valuta

SELECT name FROM Countries WHERE population<1000000
which countries have a population under 1000000
vilka länder har en befolkning under 1000000

SELECT count(*) FROM Countries WHERE population<1000000
how many countries have a population under 1000000
hur många länder har en befolkning under 1000000

SELECT Countries.name, Currencies.name FROM Countries, Currencies
WHERE continent='Europe' AND currency=Currencies.code
show the country names and currency names for all countries and
 currencies such that the continent is Europe and the currency
 is the currency code
visa landnamnen och valutanamnen för alla länder och valutor
där kontinenten är Europa och valutan är valutakoden
```

---

9 This usage is similar to T. Halpin, M. Curland, “Automated Verbalization for ORM 2”, LNCS 4278, 2006, which also suggests extensions of the CNL to serve this purpose better.

10 Technically, the abstract syntax matches relational algebra rather than SQL, and the translation first compiles SQL to algebra in the standard way. One reason is that SQL is a huge language with redundant syntactic structures that have little to do with natural language. Another is that relational algebra permits powerful transformations, normally used in query optimization but often sensible as NLG conversions as well.
As usual in GF, the translation works in all directions, thus both in query interpretation and verbalization.\footnote{The grammar is also used for generating answers, although more work is needed to structure the answers well, especially for speech interaction; cf. V. Demberg, A. Winterboer, and J. D. Moore, “A Strategy for Information Presentation in Spoken Dialogue Systems”, Computational Linguistics 37, 2011.} The query translator can also be combined with an ordinary database management system, for instance, via some Java programming using JDBC\footnote{Java Database Connectivity, \url{http://www.oracle.com/technetwork/java/javase/jdbc/}.} and GF’s Java bindings.
The Role of CNL and AMR in Scalable Abstractive Summarization for Multilingual Media Monitoring

Normunds Gruzitis and Guntis Barzdins
IMCS and LETA, University of Latvia, Riga, Latvia
normunds.gruzitis@lumii.lv, guntis.barzdins@leta.lv

In the era of Big Data and Deep Learning, a common view is that statistical and machine learning approaches are the only way to cope with the robust and scalable information extraction and summarization. Manning [1] compares Deep Learning with a tsunami at the shores of Computational Linguistics, raising a question if this is the end for the linguistically oriented approaches. Consequently, this question is relevant also to the special interest group on Controlled Natural Language (CNL).

It has been recently proposed that the CNL approach could be scaled up, building on the concept of embedded CNL [2] and, thus, allowing for CNL-based information extraction from e.g. normative or medical texts that are rather controlled by nature but still infringe the boundaries of CNL or the target formalism [3]. It has also been demonstrated that CNL can serve as an efficient and user-friendly interface for Big Data end-point querying [4, 5], or for bootstrapping robust NL interfaces [6], as well as for tailored multilingual natural language generation from the retrieved data [4].

In this position paper, we focus on the issue of multi-document storyline summarization, and generation of story highlights – a task in the Horizon 2020 Big Data project SUMMA\(^1\) (Scalable Understanding of Multilingual MediA). For this use case, the information extraction process, i.e., the semantic parsing of input texts cannot be approached by CNL: large-scale media monitoring is not limited to a particular domain, and the input sources vary from newswire texts to radio and TV transcripts to user-generated content in social networks. Robust machine learning techniques are necessary instead to map the arbitrary input sentences to their meaning representation in terms of PropBank and FrameNet [7], or the emerging Abstract Meaning Representation, AMR [8], which is based on PropBank with named entity recognition and linking via DBpedia [9]. AMR parsing has reached 67% accuracy (the F\(_1\) score) on open-domain texts, which is a level acceptable for automatic summarization [10].

Although it is arguable if CNL can be exploited to approach the robust wide-coverage semantic parsing for use cases like media monitoring, its potential becomes much more obvious in the opposite direction: generation of story highlights

---

\(^1\) http://summa-project.eu.

© Springer International Publishing Switzerland 2016
DOI: 10.1007/978-3-319-41498-0
from the summarized (pruned) AMR graphs. An example of possible input and expected output is given in Fig. 1.

| Article¹ | [...] An ongoing battle in Aleppo eventually terminated when the rebels took over the city. [...] President Assad gave a speech, denouncing the death of soldiers. [...] |
| Article² | [...] Syrian rebels took control of Aleppo. [...] |
| Article³ | [...] The Syrian opposition forces won the battle over Aleppo city. [...] Syrian president announced that such insurgence will not be tolerated. [...] |

Output Summary:

| Syrian rebels took over Aleppo | Assad gave a speech about the battle |
| Article¹ Article² Article³ | Article¹ Article³ |

Fig. 1. Abstractive summarization. An example from the SUMMA proposal

While novel methods for AMR-based abstractive\(^2\) summarization begin to appear [11], full text generation from AMR is still recognized as a future task [11], which is an unexplored niche for the CNL and grammar-based approaches.\(^3\) Here we see, for instance, Grammatical Framework, GF [12], as an excellent opportunity for implementing an AMR to text generator.

The summarized AMR graphs would have to be mapped to the abstract syntax trees (AST) in GF (see an example in Fig. 2). As GF abstract syntax can be equipped with multiple concrete syntaxes, reusing the readily available GF resource grammar library, this would allow for multilingual summary generation, even extending the SUMMA proposal.

We assume that the generation of story highlights in the open newswire domain is based on a relatively limited set of possible situations (semantic frames) and a relatively limited set of syntactic constructions (a restricted style of writing) similar to the Multilingual Headlines Generator demo\(^4\) illustrated in Fig. 3.

Although we have not yet implemented a method for the automatic mapping of AMR graphs to AST trees, there is a clear relation between the two representations. From the CNL/GF perspective, the main issue is the open lexicon (named entities and their translation equivalents), however, the AMR wiki: links to DBpedia would enable the acquisition of a large-scale multilingual GF lexicon of named entities (as implicitly illustrated in Fig. 2).

With the technique outlined in this paper, the simplified Multilingual Headlines Generator would effectively become the Multilingual Headlines Summarizer with wide applicability in the SUMMA project and beyond.

\(^2\) Abstractive summarization contrasts extractive summarization which selects representative sentences from the input documents, optionally compressing several sentences into one.

\(^3\) SemEval-2017 is expected to host a competition on AMR to text generation.

\(^4\) http://www.grammaticalframework.org/demos/multilingual_headlines.html.
Acknowledgements. This work is supported in part by the H2020 project SUMMA (under grant agreement No. 688139), and by the Latvian state research programmes NexIT and SOPHIS.

References


Author Index

Barzdins, Guntis 127
Byamugisha, Joan 25

Calafato, Aaron 1
Campo, Jesus David Nuñez 58
Cannesson, Emmanuelle 69
Cieśliński, Krzysztof 58
Colombo, Christian 1
Concepción, Eugenio 81
Condamines, Anne 69
Cuadros, Montse 37

Demirel, Ezgi 111
DeRenzi, Brian 25

Erdem, Esra 111
Etchegoyhen, Thierry 37

Fuchs, Norbert E. 13

Gervás, Pablo 81
Gruzitis, Normunds 127
Gur, Kamil Doruk 111

Jahchan, Nataly 69

Kapłański, Paweł 58
Keet, C. Maria 25
Koziółkiewicz, Jerzy 58
Kuhn, Tobias 102

León, Carlos 81
Lévy,François 92

Méndez, Gonzalo 81

Nalbandian, Christopher 48
Nazarenko, Adeline 92

Pace, Gordon J. 1
Palmaz, Susana 37

Ranta, Aarne 123

Schwitter, Rolf 48
Seganti, Alessandro 58

Wyner, Adam 92

Zarzycki, Paweł 58