View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by DCU Online Research Access Service

Revamping Question Answering with a Semantic Approach over World Knowledge

Nuno Cardoso[†], Iustin Dornescu[‡], Sven Hartrumpf[+] and Johannes Leveling[*]

[†]University of Lisbon, Faculty of Sciences, [‡]RIILP, University of Wolverhampton, UK,
[+]SEMPRIA GmbH, Düsseldorf, Germany, [*]CNGL, Dublin City University, Ireland
ncardoso@xldb.di.fc.ul.pt, I.Dornescu2@wlv.ac.uk,
hartrumpf@gmx.net, jleveling@computing.dcu.ie

Abstract. Classic textual question answering (QA) approaches that rely on statistical keyword relevance scoring without exploiting semantic content are useful to a certain extent, but are limited to questions answered by a small text excerpt. With the maturation of Wikipedia and with upcoming projects like DBpedia, we feel that nowadays QA can adopt a deeper, semantic approach to the task, where answers can be inferred using knowledge bases to overcome the limitations of textual QA approaches. In GikiCLEF, a QA-flavoured evaluation task, the best performing systems followed a semantic approach. In this paper, we present our motivations for preferring semantic approaches to QA over textual approaches, with Wikipedia serving as a raw knowledge source.

1 Introduction

Question answering (QA) is a challenging task, requiring a good expertise in several natural language processing fields to properly understand information needs in the questions, to obtain a list of answer candidates from documents, and to filter them based on solid evidence that justifies each answer's correctness.

Most state-of-the-art QA approaches are textual QA systems built around a passage retrieval core, where questions (or their affirmative re-phrasings) are treated as bag-of-words or n-grams, ignoring their semantic content. These simplified question representations are fed into the information retrieval engine to obtain the paragraphs most likely to contain answers. The candidates are then extracted based on their type and ranked based on their frequency.

Passage-retrieval QA systems have their share of success in QA evaluation tracks such as QA@CLEF, which include a considerable amount of concrete questions (such as factoid or definition questions). Answers to these questions can usually be found in the collection within a paragraph containing a sentence similar to the question (e.g. "*Where is X located ?*" → "*X is located in Y*"), or by exploiting redundant information in large collections such as the Web.

As textual QA systems focus on selecting text excerpts from the collection, they cannot address structurally complex questions that require advanced reasoning over key concepts, nor questions whose answers are not explicitly represented in a text excerpt, but must be inferred from knowledge that may (or may not) be automatically derived from those text excerpts.

GikiCLEF [1] is a QA-flavoured evaluation track that challenges participating systems to produce answers and justifications from Wikipedia for geographically-biased open list questions. A key aspect of GikiCLEF is that the answers to most topics are scattered across Wikipedia pages, and the answers and justifications have to be linked to an unambiguous real-world entity (represented by a Wikipedia URL), rather than being of pure textual nature. One of the main motivations of GikiCLEF is to refocus QA evaluation on its real challenge: *prove the answer(s) from the question, just like humans do, but automatically*; i.e., use world knowledge to reason over the entities that satisfy the criteria of the question. Although we participated in GikiCLEF with distinct systems from different projects and goals (our systems ranked 1st, 2nd, and 4th on the task [2–4]; the 3rd rank corresponds to manual experiments), we shared the same semantic vision on how to tackle the GikiCLEF task.

2 Semantic QA approach

Textual QA	Semantic QA
[web of] documents	[web of] data
document retrieval (IR core)	search for and derive facts (IE core)
query expansion on word level	question expansion on entity level
keywords & co-occurrence	concepts & relations
ambiguous words (or even word forms)	disambiguated concepts
textual snippets	graph patterns
gazetteers	RDF data
lexical semantics (thesaurus oriented)	formal semantics
text with entities	**linked entities with text**

Table 1. Key differences between textual QA and semantic QA approaches

Table 1 summarizes what we think are the key differences of textual QA approaches, compared to semantic QA approaches like ours. Textual QA is typically based on retrieving candidate answers from textual snippets in a document collection, augmented by term-based statistical approaches such as query expansion on the level of words or conflated word forms (stems). In contrast, semantic QA employs processes that go *beyond* matching questions and documents: external knowledge in a formal representation (such as RDF) is used to reason over disambiguated concepts and entities, derive relations between them, and infer answers from the question representation. While textual QA approaches can be successful for finding explicitly stated answers from documents, semantic QA aims for complex questions where several information sources must be merged.

2.1 Accessing the world knowledge

One of the obstacles for semantic QA approaches is the need for a large repository of "condensed world knowledge" derived from validated, well-structured and machine-accessible data to enable inferring information that is implicit in

the text. While compiling such a repository is highly unfeasible for any single researcher or research group (the famous knowledge bottleneck), it is feasible for community-driven projects such as Wikipedia or Freebase. Along with upcoming projects like DBpedia [5], YAGO [6] and the recent interest in Linked Data [7], the QA community now has at its disposal large amounts of human knowledge in machine-readable format that is easily accessible, freely available, and constantly improving in quantity and quality. While such resources and services only cover certain domains and do not (yet) encompass all the knowledge required to answer all questions from either QA evaluation fora or from real users, the foundation is now laid and we can start developing QA systems with semantic strategies over this Linked Data layer of human knowledge.

2.2 From question analysis to answer reasoning

Our semantic QA systems dedicated special attention to these three tasks:

Grounding expected answer types to a workable category/classification. The expected answer type (EAT) for a question must be mapped to *entities* from the world knowledge, so that its meaning becomes unambiguously grounded. Take for instance the question "*Which Romanian writers were born in Bucharest in the 20th century?*"; the EAT can be grounded to DBpedia resource URL `http://dbpedia.org/resource/Category:Romanian_writers` which makes it easier to search and validate candidate answers.

Parse constraints from the question. A semantic question interpretation can be seen as an EAT and a composition of constraints. In the example above, the question has two restrictions: one geographical (*born 'in Bucharest'*) and one temporal (*born 'in the 20th century'*). By handling each constraint separately, systems can use its meaning to employ the right strategies for validating candidate answers [3]. This divide-and-conquer method is well illustrated by Cardoso et al. [2], who mapped predicates in the conditions into DBpedia ontology properties, or by Hartrumpf et al. [4] who decompose a question into subquestions and uses subanswers to reformulate the original question.

Semantic analysis. The natural language question is transformed into a machine-readable representation, which can be rewritten or expanded in some way. One technique is logical question expansion by employing an inference engine with rules for paraphrases, entailments, and logical equivalence, as described in [4]. Another technique is to generate SPARQL queries from all grounded concepts in the question, and submit to DBpedia's SPARQL endpoint (`http://www.dbpedia.org/sparql`), as shown in [2]. Relying solely on knowledge bases such as DBpedia to obtain answers leads to very high precision at the cost of very low recall, but this will change as Wikipedia articles become more complete, and DBpedia's ontology coverage on extracted resources and grounded properties increases. A reference such as `http://dbpedia.org/resource/Mihail_Fărcăşanu` is not just a span of text of type *Person* – as is the case for textual QA, but

an entity linked to various KB, both structured (databases, ontologies) and unstructured (textual). Disambiguated references allow merging relevant information sources as well as combining heterogeneous procedures. It is this background information that can help answer questions traditionally considered "complex".

3 Final Remarks

The QA community finally has access to a considerable amount of world knowledge, created and curated by a dedicated community, encoded in machine-readable formats such as RDF/OWL and freely available through Linked Data. We feel that the time has come to focus on semantic approaches for QA. The core strategy of semantic QA systems is to *understand* key entities, concepts and relations between them in the question (the information need), and solve it by exploiting multiple knowledge sources and selecting the best strategies to determine and validate the right answers. Classic textual QA approaches are still useful for simple questions, but are quite limited for elaborated questions, and depending on a collection limits their question range. We foresee that the performance of semantic QA systems will improve as the resources and services from associated projects evolve over time, and will be able to answer more complex, open list questions that have more resemblance to real-user information needs.

Acknowledgments

This work is partially supported by FCT for its LASIGE Multi-annual support, GREASE-II project (grant PTDC/EIA/73614/2006) and Nuno Cardoso's scholarship (grant SFRH/BD/45480/2008), partly supported by the EU-funded project QALL-ME (FP6 IST-033860), and supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (CNGL).

References

1. Santos, D., Cabral, L.M.: GikiCLEF: Crosscultural issues in an international setting: asking non-English-centered questions to Wikipedia. In: Working Notes for CLEF 2009, Corfu, Greece (30 Sept.-2 Oct. 2009)
2. Cardoso, N., Baptista, D., Lopez-Pellicer, F.J., Silva, M.J.: Where in the Wikipedia is that answer? the XLDB at the GikiCLEF 2009 task. In: Working Notes for CLEF 2009, Corfu, Greece (30 Sept.-2 Oct. 2009)
3. Dornescu, I.: EQUAL: Encyclopaedic QUestion Answering for Lists. In: Working Notes for CLEF 2009, Corfu, Greece (30 Sept.-2 Oct. 2009)
4. Hartrumpf, S., Leveling, J.: GIRSA-WP at GikiCLEF: Integration of Structured Information and Decomposition of Questions. In: Working Notes for CLEF 2009, Corfu, Greece (30 Sept.-2 Oct. 2009)
5. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: A Nucleus for a Web of Open Data. In: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ASWC 2007, Busan, Korea, Nov. 11–15, 2007. Number 4825 in LNCS. Springer (2007) 722–735
6. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: A Core of Semantic Knowledge. In: Proceedings of the WWW'07, Banff, Canada, ACM (May 8-12 2007)
7. Berners-Lee, T.: Linked Data - Design Issues. `http://www.w3.org/DesignIssues/LinkedData.html` (2007)