Computational Semantics and Information Retrieval

Christof Monz
Institute for Logic, Language and Computation (ILLC)
University of Amsterdam, Plantage Muidergracht 24,
1018 TV Amsterdam, The Netherlands
christof@wins.uva.nl

1 Introduction

In this talk I want to review the area of Information Retrieval (IR) from a computational semanticist's point of view. There have been several attempts to integrate Natural Language Processing (NLP) techniques into IR and I want to discuss some of those attempts and point out some opportunities and challenges for computational semantics.

Research in IR is aimed at designing and evaluating systems that try to fulfill a user's information need. Often, this is a user posing a query to a database. An IR system can react in a variety of ways. For instance, it can present a list of documents presumably containing the information the user is looking for (ad-hoc retrieval), or it can directly answer a user's question by generating natural language sentences or extracting sentences from the database (Question-Answering). The database itself is simply a collection of natural language documents. From an abstract point of view, a document d is relevant to a query q if d 'is about' q. Defining 'is about' is a non-trivial task and several approaches have been proposed, cf. Baeza-Yates and Ribeiro-Neto (1999) for an overview.

One of the hallmarks of the field of IR is its emphasis on testing and evaluation. Experiments are usually carried out on standard collections, such as the TREC collections (Harman, 1995), where the relevant documents for a query are known in advance. To evaluate the quality of an IR system standard measures such as *precision* and *recall* are used. Precision compares the number of relevant documents that have been retrieved to the number of non-relevant documents. Recall indicates the number of relevant documents that have been retrieved to the number of documents which are considered to be relevant.

2 Semantic Representation

Usually, information retrieval systems do not work on the documents themselves, but on a (semantic) representations. Therefore, deciding whether a document is relevant to a query depends on the kind of representation that is being used for the document and for the query. Almost all existing IR systems simply represent documents and queries as a 'bag-of-words'. From a formal semanticist's point of view this may seem hopelessly inadequate, but for simple retrieval tasks such as ad-hoc retrieval this way of representing the content of a document turns out to be surprisingly effective. It may seem intuitively obvious that exploiting linguistic structure will help to improve retrieval effectiveness, and that deeper or richer representations, although perhaps more computationally costly, will lead to a substantially higher precision. Indeed, there have been several attempts to add multi-word phrases to document representations, see e.g. Strzalkowski (1995), Mitra et al. (1997), but the experimental findings on retrieval with such enhanced representations do not really

support the hypothesis that added linguistics structure improves effectiveness. In some cases, it indeed does improve retrieval, but the results are not uniform enough to speak of a significant improvement.

Despite its success, there are several problems tied to the simple bag-of-words' presentation of documents. Two of these problems are of particular relevance to computational semanticists: word-sense ambiguity and synonymy. If a query contains a word which is lexically ambiguous, it may happen that documents are retrieved which contain this word, but not in its intended meaning. Conversely, it may happen that a document is not retrieved because it does not share a word with the query, although it does contain words which are synonymous to words in the query.

Most approaches to word-sense disambiguation or to finding synonyms employ word taxonomies like WORDNET (Miller, 1995). To disambiguate word-senses a word is tagged with one of its senses, for instance a WORDNET synset, induced by the context of the occurrence of the word, cf. Sanderson (2000). And synonyms can be exploited by a technique called query expansion, where synonymous terms are added to the query terms, cf. Voorhees (1994). As in the case of using multi-word phrases, experimental findings do not indicate a significant improvement in effectiveness when retrieving with disambiguated word senses or synonyms, see Voorhees (1994). Those experimental results suggest that employing NLP techniques for extracting linguistic structures and using them for ad-hoc retrieval will not significantly improve retrieval effectiveness.

It could be argued that the failure of NLP techniques here is simply due to the very shallow and limited character of ad-hoc retrieval. Maybe, less shallow information needs that require a deeper analysis of the documents and queries can profit from the use of NLP techniques. A prime example here are Question-Answering systems; see Kupiec (1993) and Voorhees and Tice (1999). Such systems do not return full documents but return a single (partial) sentence which is supposed to be an answer to the user's input question. To get an impression what typical questions look like consider examples (1–3), which are questions that have been posed to ASKJEEVES (ASKJEEVES, 2000), an on-line Question-Answering system.

(1) What movie took the longest to film?

(2) What is a mashuganas?

(3) How many members of the U.S. Congress and Senate are graduates of Bob Jones University? And what are there names?

In contrast to ad-hoc retrieval, Question-Answering is a task that seems to be much more sensitive to linguistic structures like multi-word phrases and argument structure, thus potentially raising a whole series of interesting challenges for NLP in general and computational semanticists in particular.

To focus on the latter, one of the prime issues that has to be addressed is: How do we represent argument structure? One way to represent argument structures is to use *description logic* formulas. Description logics are a family of knowledge representation languages originating from work on semantic networks and frame-based formalisms, see Donini et al. (1996). For reasons of robustness and efficiency, partial parsing techniques are used to construct description logic representations. While this makes it impossible to provide a deep semantic representation, for most information retrieval tasks, including Question-Answering, this is not necessary anyway. Since typical wh-questions are of the form "Who did what to whom and when?", the information that we do need to capture in the representation includes thematic information, such as which NP is the agent and which one is patient etc., see Litkowski (1999). Description logical formulas capturing this information can look like (4) and (5).

(4) a. Industry sources put the value of the acquisition at $100 million

b. $\exists \texttt{agent}.(\texttt{source} \sqcap \texttt{industry}) \sqcap \exists \texttt{event}.\texttt{put} \sqcap \exists \texttt{patient}.(\texttt{acquisition} \sqcap \exists \texttt{value}.(100,000,000 \sqcap \texttt{dollar}))$

(5) a. John Blair was acquired last year by Reliance Capital Group Inc.

b. $\exists \texttt{agent}.\texttt{reliance\_capital\_group\_inc} \sqcap \exists \texttt{event}.\texttt{acquire} \sqcap \exists \texttt{patient}.\texttt{john\_blair} \sqcap \exists \texttt{time}.\texttt{last\_year}$

This way of using description logic to provide only a shallow semantic representation is based on Meghini et al. (1993).

3 Inference

Once we have representations, we can perform inferences with them. What kinds of inference tasks these will be, will depend both on the representations themselves and on the information need. In the case of ad-hoc retrieval, the representations are bags of words and the corresponding inference task is simply term matching.

As we move to retrieval tasks that require deeper representations, such as Question-Answering, the corresponding reasoning tasks become more complex. In Question-Answering, the user's information request is more specific than in ad-hoc retrieval. For instance, wh-questions ask for information concerning a particular argument position, and a Question-Answering system has to be able to identify argument positions in the documents to answer the question properly. In the context of wh-questions, inference amounts to comparing the argument structures in the documents to the argument structures of the query and in case of a match returning the value of the wh-argument.

The simplest way of performing inference is to use template matching, where a template is an argument structure and the wh-argument is left empty. A shortcoming of this approach is that perfect matches can be expected to be rather rare. The values of arguments can be complex phrases and it is necessary to split the phrases into their simpler constituents such as head-modifier pairs. Comparing only the heads increases the chance of matching.

But even then, we are still facing problems like word-sense ambiguity and synonymous words — just as in ad-hoc retrieval. Given the quite disappointing experimental results of applying word-sense disambiguation and query expansion with synonyms to ad-hoc retrieval, the obvious question is whether comparable results can be expected if these techniques are applied to Question-Answering. The answer is unclear as this has not been carefully investigated so far. Similarly, if two argument values do not match, but one is a hyponym of the other, should this be considered as a match?

It seems clear that we have to investigate the opportunities of hierarchical inferences further. This supports the use of description logic for representing argument structures because it allows for a more flexible manipulation than templates. To manipulate the proposed semantic representations (description logic formulas), many highly optimized high-quality tools such as FaCT (Horrocks, 1999) and RACE (Haarslev and Möller, 1999) are available. Using description logic for manipulating semantic representations has the added advantage that it can be easily integrated with hierarchical reasoning tasks, because drawing hierarchical inferences is at the very heart of description logic.

4 Conclusions

While I think that there is a role to be played by computational semantics in information retrieval, special attention should be paid to the balance between suitable representations and corresponding inference tasks on the one hand, and the retrieval tasks for which they are being put to use on the other hand.

This position should be contrasted with logic-related work in information retrieval that was carried out in the early days of information retrieval. Van Rijsbergen (1986) defined

'about-ness' of a document d in terms of logical entailment, where d 'is about' q if $d \models q$. Logical entailment could then be checked automatically by using theorem proving techniques. Thus, to retrieve the documents relevant for a query q, one has to check whether $d \vdash q$ for each document d in the collection. Although there are efficient theorem provers like Bliksem (de Nivelle, 2000) and SPASS (Weidenbach et al., 1996), applying them to standard data collections like TREC (Harman, 1995) which consist of several hundreds of thousands of documents, this turns out to be computationally very challenging, cf. Crestani et al. (1995). The latter illustrates once again the importance of experimental testing in IR — an important methodological hallmark which should prove instructive or even refreshing for Computational Semantics

Acknowledgments. I want to thank Maarten de Rijke for his helpful comments. The author was supported by the Physical Sciences Council with financial support from the Netherlands Organization for Scientific Research (NWO), project 612-13-001.

References

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison Wesley.

Crestani, F., Ruthven, I., Sanderson, M., and van Rijsbergen, C. (1995). The troubles with using a logical model of IR on a large collection of documents. In Harman, D., editor, *Proceedings of the 4th Text Retrieval Conference (TREC-4)*, pages 509–526. NIST Special Publication 500-236.

de Nivelle, H. (Accessed July 2000). Bliksem resolution prover.
`http://www.mpi-sb.mpg.de/~bliksem/`

Donini, F., Lenzerini, M., Nardi, D., and Schaerf, A. (1996). Reasoning in description logics. In Brewka, G., editor, *Principles of Knowledge Representation*, pages 191–236. CSLI Publications.

Haarslev, V. and Möller, R. (1999). RACE system description. In Lambrix, P., Borgida, A., Lenzerini, M., Möller, R., and Patel-Schneider, P., editors, *Proceedings o the International Workshop on Description Logic (DL99)*, pages 130–132.

Harman, D. (1995). The second text retrieval conference (TREC-2). *Information Processing & Management*, 31(3):271–289.

Horrocks, I. (1999). FaCT and iFaCT. In *Proceedings o the International Workshop on Description Logic (DL99)*, pages 133–135.

Kupiec, J. (1993). MURAX: A robust linguistic approach for question answering using an on-line encyclopedia. In *Proceedings of the 16th annual international ACM SIGIR conference*, pages 181–190.

Litkowski, K. (1999). Question-answering using semantic relation triples. In *Proceedings of the 8th Text Retrieval Conference (TREC-8)*.

Meghini, C., Sebastiani, F., Straccia, U., and Thanos, C. (1993). A model of information retrieval based on a terminological logic. In Korfhage et al., R., editor, *Proceedings of SIGIR-93*, pages 298–307. ACM Press, Baltimore.

Miller, G. (1995). WORDNET: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Mitra, M., Buckley, C., Singhal, A., and Cardie, C. (1997). An analysis of statistical and syntactic phrases. In *Conference Proceedings of RIAO-97*, pages 200–214.

Sanderson, M. (2000). Retrieving with good sense. *Information Retrieval*, 2(1):49–69.

Strzalkowski, T. (1995). Natural language information retrieval. *Information Processing & Management*, 31(3):397–417.

ASKJEEVES (Accessed July 2000). `http://www.aj.com`

van Rijsbergen, K. (1986). A non-classical logic for information retrieval. *The Computer Journal*, 29(6):481–485.

Voorhees, E. (1994). Query expansion using lexical-semantic relations. In *Proceedings of the seventeenth annual international ACM-SIGIR conference on Research and development in information retrieval*, pages 61–69.

Voorhees, E. and Tice, D. (1999). The TREC-8 question answering track evaluation. In *Proceedings of the 8th Text Retrieval Conference (TREC-8)*.

Weidenbach, C., Gaede, B., and Rock, G. (1996). SPASS & FLOTTER, version 0.42. In *13th International Conference on Automated Deduction, CADE-13*, LNAI. Springer.