The Knowledge Engineering Review, Vol. 17:1, 71–75. © 2002, Cambridge University Press
DOI: 10.1017/S026988890200036X Printed in the United Kingdom

Ontologies and text retrieval*

JAMES MAYFIELD

The Johns Hopkins University Applied Physics Laboratory, 11100 Johns Hopkins Rd., Laurel MD 20723–6099, USA; e-mail: James.Mayfield@jhuapl.edu

1 Introduction

Analogues to much of today's work in ontologies have existed for centuries in text retrieval. The use of controlled vocabularies, or *thesauri*, has been fundamental to document indexing in library science. Thesauri serve several purposes, including:

- **Knowledge organisation** A thesaurus provides a hierarchy of concepts that organises domain-specific knowledge.
- **Terminology normalisation** By selecting a unique word or phrase to represent each domain concept, then linking synonymous terms to it, a thesaurus enforces terminological consistency.
- **Query expansion** A thesaurus facilitates the addition of terms to a query by providing explicit hierarchical and lateral relationships among terms.

These properties serve to mediate the information flow from indexer to user. Thesauri thus serve many of the same functions for people that ontologies are designed to serve for software agents. As automated retrieval has developed over the decades since the inception of computer processing of text, many techniques have been introduced to apply this typically manual work to the automated arena (see Soergel (1985) for an introduction to library information systems, also Anderson and Pérez-Carballo (2001a, 2001b) for a summary of the intersection of human and machine indexing).

In this article, I describe the impact of work in machine-readable ontologies on automated text retrieval. I focus on two areas that are fundamentally new to text retrieval with the advent of machine-readable ontologies: the reduced granularity of passages that may be conveniently described by the ontology and the existence of large amounts of parallel ontological/textual data. I then describe outstanding problems that must be solved for the techniques outlined herein to come to fruition. I will restrict my comments to text retrieval, because it is the best-studied subfield in information retrieval (IR), and has the highest potential for contribution to work in ontologies. However, many of my comments will apply to other types of retrieval as well, such as audio, video or image retrieval. Also note that for brevity I typically blur the distinction between the elements of an ontology proper, which represent concepts, and instantiations of those concepts.

2 Explicit component-level semantics

Because it has typically been done manually, the use of thesauri to describe document content has almost always been applied at the level of large units of text such as books or articles. In contrast, automated indexing methods have typically used the entire contents of a document (e.g. all of its

* This work was supported in part by the Defense Advanced Research Projects Agency (DARPA) under contract number F30602-00-2-0 591 AO K528. I would like to thank Tim Finin, who was instrumental in my involvement in this line of enquiry.

words) as indexing terms for that document, but without the strong semantics conveyed by thesaurus entries. The promise of ontologies for text retrieval is that they may allow strong semantics to be applied to the individual paragraphs, sentences and even words of documents to be indexed.

Assuming that appropriate ontological representations of a text can be derived (a topic discussed below under "Outstanding problems"), the most obvious way to associate those representations with the original text is with markup. SGML has provided a widely accepted mechanism for encoding markup of various sorts in otherwise unstructured text. Even before the widespread adoption of HTML, numerous SGML standards were springing up in many fields. While HTML conveys only syntactic information, the Web is now catching up through a variety of standards that capture increasing amounts of semantics in markup. The DAML + OIL specification[1] is the de facto standard for encoding ontologies as markup, and the Semantic Web Activity[2] of the World Wide Web Consortium[3] will likely produce a standard that is widely accepted.

The presence of low-level semantic markup in text makes possible detailed matching of query semantics against document semantics. For example, a query about the bestowers of Nobel prizes can avoid retrieving documents about Nobel prize recipients by matching semantic structures. Nor is matching restricted to the semantic forms explicitly present in the query and document – any inference process that derives new semantic structures might be applied to query or document before matching proceeds.[4] The accuracy that might be achieved in matching semantic structures is quite beyond anything available today using statistical methods, and will be crucial for exploitation of text by agents that lack human mediators.

3 Association between text and semantic structures

The existence of large amounts of text augmented with semantic markup makes it possible to exploit the statistics of such collections to find relationships between text and markup. A *parallel collection* is a bilingual document[5] collection in which each document written in one language has a translation into the other language. Parallel texts are useful for cross-language retrieval, in which a query in one language must select relevant documents in another language. A typical approach is to search for a query term in the source language collection, then select (using mutual information, co-occurrence statistics and so on) words that characterise the corresponding target language documents; the resulting "translations" serve as a target language query, which is then put against the actual collection of interest (see Yang *et al*. (1998) and Nie *et al*. (1999) for examples of this approach).

A text containing semantic markup is effectively a parallel text between human language and semantic structures. Thus techniques for exploiting cross-language parallel collections can be applied to collections that contain both text and ontological markup; we simply treat text and ontology as two different languages, and use the collection to translate between them. Two complementary techniques are then available. First, we can find examples of specific semantic markup (or identify entire ontologies) that might be applicable to a new text selection by using the parallel collection to map from the text to semantic structures. This might be valuable, for instance, to a person marking semantic structures by hand for later consumption by an agent, or to a software agent searching for a suitable published ontology. Second, we might use the same process in the other direction to find text that describes semantic structures. That is, we search for documents containing semantic structures that are similar to our input structure, then extract important words from the associated text. This technique might help, for example, with finding nodes in two different ontologies that express the same concept (see **ontology alignment** below) – two nodes are candidates for pairing if the words that describe them are similar.

[1] `http://www.daml.org/2001/03/daml+oil-index.html`.
[2] `http://www.w3.org/2001/sw/`.
[3] `http://www.w3.org/`.
[4] While control of inference and matching presents many difficulties, these difficulties are rarely IR-specific – they arise in any domain where the potential inferences are exponential in the number of extant facts and rules.
[5] "Document" is used here as a generic term; it may be as short as a sentence or phrase.

4 Outstanding problems

Three outstanding problems are most significant relative to the interaction between ontologies and text retrieval.

Markup: *how do we correctly tie large amounts of text to specific ontologies?* In the limit this problem reduces to the long-studied, extremely difficult one of general Natural Language Understanding (NLU). However, there are many techniques shy of full NLU that show promise for automatically creating ontological markup. Named entity recognisers (many developed for the Message Understanding Conferences[6]) use finite state recognisers or hidden Markov models to identify noun phrases that represent the major actors in a text. Work in word-sense identification (mapping word uses onto a known set of word senses) and induction (clustering word uses to form derived senses) suggests how words and phrases might be automatically related to the nodes of an ontology (Schütze & Pedersen, 1995; Yarowsky, 2000). Attempts to exploit word-sense identification for text retrieval have produced many negative findings and relatively few positive ones. Sanderson (2000) observes that word-sense assignment must be highly accurate to convey benefit to text retrieval, a requirement that is likely to be shared by agents attempting to use derived semantics. Specific semantic resources, most notably WordNet (Fellbaum, 1998) and Cyc (Lenat & Guha, 1990), have served as basic resources for text interpretation. WordNet has also been used for automated word-sense assignment (Gonzalo *et al*., 1998; Mihalcea & Moldovan, 2000). Attempts to automate the creation of semantic resources such as WordNet (Grefenstette, 1994; Hearst, 1998) typically meld corpus-based techniques with linguistic analysis. Finally, question-answering systems (e.g. FALCON (Harabagiu *et al*., 2001)) roll together many of these technologies to identify basic subject-verb-object relationships that might represent answers to particular questions. Despite these many advances, automatic identification of semantic relationships in text remains the largest roadblock to agent exploitation of text.

Tagging a text relative to a particular ontology restricts the range of entities and relationships that must be considered during tagging to those representable by the ontology, so the difficulty of the markup problem can be reduced by restricting the ontology.

A related problem deals with markup heterogeneity: *how do we produce markup that doesn't look like it fell out of a database?* Early efforts to develop a Semantic Web (Berners-Lee *et al*., 2001) have produced significant quantities of ontological "markup".[7] However, much of this markup has an extremely regular structure, because it is created by exploiting boilerplate texts and structured databases. For example, the CIA World Fact Book,[8] which provides geographic, political and economic information about the countries of the world, is amenable to ontological markup without significant NLU technology because of the regularity of its entries. However, the resulting markup displays little variability. Statistical techniques that rely on differences in term frequency fail when all terms have the same frequency. Thus, while markup homogeneity may be beneficial to an agent that is simply trying to access the facts represented in markup, it will inhibit the agent that tries to bridge statistically the worlds of formal semantics and text. This problem will be resolved only as more sophisticated markup techniques are put into common use. While this will eventually occur, the speed with which they are adopted will affect the perceived usefulness of pairing ontologies with text-retrieval techniques.

Ontology alignment: *how do we align two ontologies?* When two texts are marked up relative to different ontologies, we can only draw meaningful conclusions about the semantic relationships between the texts if we can characterise the relationships between the nodes of the two ontologies. Of course, all of the extant difficulties in ontology manipulation affect the area of intersection between ontologies and text retrieval. However, ontology alignment is particularly important for text retrieval because the promise of ontologies in text retrieval is that they allow querying at the semantic level where ordinary text retrieval is confined to the lexical level. We cannot expect all texts to be marked

[6] `http://www.itl.nist.gov/iaui/894.02/related\_prrojects/muc/`.
[7] `http://www.daml.org/crawler/pages.html`.
[8] `http://www.cia.gov/cia/publications/factbook/`.

relative to a single universal ontology. Therefore, to search over collections containing markup that refers to more than one ontology, we will need the ability to map concepts expressed in one ontology onto the equivalent concepts in other ontologies.

Querying: *how do we specify queries that have both unstructured and structured components, and match them against documents that contain text, semantic markup or both?* People are accustomed to generating unstructured queries to search text collections, but typically have little experience generating structured queries. Software agents, on the other hand, excel at generating structured queries but have had no standard, integrated way to exploit unstructured textual data. Some structured query languages, such as variants of SQL, allow limited specification of unstructured text; however, the matching that is done on such specifications is usually restricted to exact substring match within a single text field of the database. A hybrid query language for use with ontologies must allow text and semantic structure to be interspersed. Document ranking has proven invaluable in text search; to allow ranking of results for hybrid queries, partial matching of semantic structures must be supported. Finally, because the types of processing that might be applied during matching of semantic structures are so wide-ranging, it is reasonable to expect that the query language would allow explicit control over such processing. These requirements suggest that while software agents may have no difficulty generating appropriate hybrid queries, people may well require user interfaces or personal agents that hide some of the query details. Many proposals for query languages were presented at the 1998 Query Languages Workshop;[9] standardisation is under discussion by the W3C XML Query working group.[10]

5 Conclusions

The challenge in automated text retrieval continues to be "beating the DOW". That is, much as an approach to stock purchase is not worth much if it cannot outperform the simple strategy of purchasing uniformly from a fixed set of stocks (such as those that make up the Dow Jones Industrial Average), so an approach to text retrieval that cannot outperform simple statistical approaches is of questionable value. Ontologies hold the potential to improve significantly the information search experience for users (be they people or software agents), because they explicitly and unambiguously encode information that is typically implicit and ambiguous in text. However, this will only occur if solutions to the problems outlined above are found. In the interim, the dual nature of the relationship between ontologies and text will allow text retrieval techniques to provide a needed boost to the development of ontologies and their subsequent exploitation by software agents.

References

Anderson, JD and Pérez-Carballo, J, 2001a, "The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing" *Information Processing and Management* **37**(2) 231–254.

Anderson, JD and Pérez-Carballo, J, 2001b, "The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part II: Machine indexing, and the allocation of human versus machine effort" *Information Processing and Management* **37**(2) 231–254.

Berners-Lee, T, Hendler, J and Lassila, O, 2001, "The semantic web" *Scientific American* **284**(5) 35–43.

Fellbaum, C (ed.), 1998, *WordNet: An Electronic Lexical Database* MIT Press.

Gonzalo, J, Verdejo, F, Chugur, I and Cigarrán, J, 1998, "Indexing with WordNet synsets can improve text retrieval" *Proceedings of the COLING/ACL '98 Workshop on Usage of WordNet for NLP* 38–44.

Grefenstette, G, 1994. *Explorations in Automatic Thesaurus Discovery* Kluwer Academic Publishers.

Harabagiu, S, Moldovan, D, Paşca, M, Mihalcea, R, Surdeanu, M, Bunescu, R, G'rju, R, Rus, V and Morărescu, P, 2001, "FALCON: boosting knowledge for answer engines" *Proceedings of the Ninth Text REtrieval Conference* (TREC-9) 479–488. Also available from `http://trec.nist.gov/pubs/trec9/t9\_proceedings.html`.

[9] `http://www.w3.org/TandS/QL/QL98/`.
[10] `http://www.w3.org/XML/Query`.

Hearst, MA, 1998, "Automated discovery of WordNet relations" in C Fellbaum (ed.) *WordNet: An Electronic Lexical Database* MIT Press.

Lenat, D and Guha, RV, 1990, *Building Large Knowledge-Based Systems* Addison-Wesley.

Mihalcea, R and Moldovan, D, 2000, "Semantic indexing using WordNet senses" *Proceedings of the ACL 2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval* pp-pp. Also available from `http://sensei.ieec.uned.es/IRNLP-2000/papers/`.

Nie, J-Y, Simard, M, Isabelle, P and Durand, R, 1999, "Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web" *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '99) 74–81.

Sanderson, M, 2000, "Retrieving with good sense" *Information Retrieval* **2**(1) 47–67.

Schütze, H and Pedersen, JO, 1995, "Information retrieval based on word senses" *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval* 161–175.

Soergel, D, 1985, *Organizing Information: Principles of Data Base and Retrieval Systems* Academic Press.

Yang, Y, Carbonell, JG, Brown, RD and Frederking, RE, 1998, "Translingual information retrieval: learning from bilingual corpora" *Artificial Intelligence* **103**(1–2) 323–345.

Yarowsky, D, 2000, "Word sense disambiguation" in R Dale, H Moisl and H Somers (eds.) *The Handbook of Natural Language Processing* Marcel Dekker.