Chapter 4
NLP for Search

Christian F. Hempelmann
Purdue University, USA

ABSTRACT

This chapter presents an account of key NLP issues in search, sketches current solutions, and then outlines in detail an approach for deep-meaning representation, ontological semantic technology (OST), for a specific, complex NLP application: a meaning-based search engine. The aim is to provide a general overview on NLP and search, ignoring non-NLP issues and solutions, and to show how OST, as an example of a semantic approach, is implemented for search. OST parses natural language text and transposes it into a representation of its meaning, structured around events and their participants as mentioned in the text and as known from the OST resources. Queries can then be matched to this meaning representation in anticipation of any of the permutations in which they can surface in text. These permutations centrally include overspecification (e.g., not listing all synonyms, which non-semantic search engines require their users to do) and, more importantly, underspecification (as language does in principle). For the latter case, ambiguity can only be reduced by giving the search engine what humans use for disambiguation, namely knowledge of the world as represented in an ontology.

INTRODUCTION

This chapter could have been written as an intro to applying standard Information Retrieval (IR) techniques to internet search as these techniques are the basis for most approaches to search today (“have method, looking for application”). In a nutshell, IR techniques operate by identifying desired keywords or their clusters in a collection of texts and retrieving document, for example www pages, that match the keywords. But such introductions have been done elsewhere and better than this author could. This chapter could also have been written as a theoretical comparison of IR and Information Extraction (IE) techniques, based on the tenets of research in Artificial Intel-
lingence (AI), which is where NLP contributions to search seem to be headed, as I will argue. To put it simply, IE techniques aim to ‘understand’ text to varying degrees and extract the relevant small bits in relation to the information needs of users (for a generic system, see Hobbs 1993). But such introductions have also been done elsewhere and better than this author could. Instead this chapter is going to sketch these issues in its introduction, before focusing on one application in this new direction, based on the experience of its author, namely in building and improving a search engine that is based on representation of meaning with the help of linguistic AI and facilitating IE-style search. As such, this chapter will largely ignore non-linguistic problems and solutions in search.

As for the majority of areas in NLP, text search is largely dominated by statistical approaches. The basic issue for any ANLP is that the complexity, some call it mess, of natural language needs to be made palatable to the computer, to discover in or impose on the unstructured mess of language some formal structure. This formal representation of language, and hopefully some aspect of its meaning gleaned from its surface structure, can then be used by the computer with any formal algorithm, hopefully suggested by a theory for a given application, but often just the favored algorithm of a research group in search for new applications.

Such non-linguistic approaches choose to ignore that language is language and operate under the assumption that its surface manifestation, in particular co-occurrence in its surface representation, are a sufficient window on the underlying meaning. After all, meaning is what all approaches are after, because it is the level at which humans interface with each other through language, and the meaning of language does indeed correlate with its surface manifestation, the text, to a large degree. But the degree to which meaning doesn’t surface repeatedly and regularly in natural language text is inaccessible to statistical methods and responsible for there being an ultimate limit to what these methods can achieve. Furthermore, “language events” are very sparse, which can be gleaned from the famous observation that in a large corpus, trigrams are 85% unique (which can be alleviated to some degree through smoothing and extraction). In other words, of the sequences of three words in a text, the large majority does not recur.

Another approach, actually the rationale of IR in contrast to IE, is to assume that ultimately humans will be the consumers of the application’s output. Under this assumption, human searchers are sufficiently served by documents to his or her query that are deemed relevant by the computer because of overlap to the query and other relatively easily formalized ranking factors. The humans can then extract the information from those documents that fill their information needs on their own, that is, the machine doesn’t have to do semantics, since the human is at the end of the processing chain. In contrast to this, the assumption in the main part of this chapter is that giving the machine semantics to use in matching and ranking will improve its performance and decrease human work load, both common main motivations in automation. In sum, on the basis of the unit concept, the computer represents the meaning of documents and fills the information need of the human from a knowledge base, not a document base (cf. Spärck Jones 1990).

**ADDING NLP TO SEARCH**

**To Know or to Guess**

In the field of language and computers, the fashion changes back and forth between representational and statistical dominance, under different names and with different ‘applications’ on the banners of the proponents of either side. As mentioned, NLP for Internet search is largely identical with the methods traditionally described under the term Information Retrieval (IR). Recently, and not