Searching and Search Engines:
When is Current Research Going to Lead to Major Progress?

Elizabeth D. Liddy
Professor, School of Information Studies
Director, Center for Natural Language Processing
Syracuse University
New York, USA

Abstract

For many years, users of commercial search engines have been hearing how the latest in information and computer science research is going to improve the quality of the engines they rely on for fulfilling their daily information needs. However, despite what is heard, these promises have not been fulfilled. While the Internet has dramatically increased the amount of information to which users now have access, the key issue appears to be unresolved – the results for substantive queries are not improving. However, the past need not predict the future because sophisticated advances in Natural Language Processing (NLP) have, in fact, accomplished significant improvements in engines that can provide ease of access for users as well as improved quality of retrieved information.

Introduction

Search engines provide the interface between queries and documents, both of which are language artifacts. Therefore, it makes sense for search engines - whose task is to find documents (or web pages or answers) in response to users' queries - to utilize natural-language-based processing and representation to accomplish the task. However, search engines have not historically utilized real language processing – but instead have either relied on simple string matching or statistical processing of text. This apparent avoidance and only late adoption of NLP by commercial search engines raises many questions that need to be answered – Have NLP approaches been implemented? How was it done? How well did it perform? Was it done adequately & appropriately? If not, is there reason to believe that the state of the art of NLP research is currently at a stage that we can expect major progress from utilizing NLP?

To answer these questions, this paper will provide a brief introduction to NLP; provide some indication of how current search engines do or do not use NLP; present the real state of the art in NLP when done appropriately; describe applications which fully utilize the capabilities of NLP in information access tasks of interest to this audience, including Information Retrieval, Information Extraction, and Question-Answering, and finally; conclude with a statement of how and when these NLP-based improvements will be available for your benefit.

NLP Primer

Natural Language Processing is an approach (comprised of both theory and technology) that enables a system to accomplish human-like processing and interpretation of text by its ability to recognize and represent the implicit as well as the explicit meaning in a text by utilizing all the levels of human language understanding. This includes the morphological, lexical, syntactic, semantic, discourse, and pragmatic levels of language (see Figure 1). The goal of NLP-based technologies is to simulate human extraction of meaning from all of these levels of language. As explained in more detail in Liddy, (1998), each of these levels of language conveys a different type of meaning. And while we frequently hear the use of the word 'semantic' as if it equated to 'meaning' - in fact, all the levels of language convey meaning to human understanders. Moving up from morphology, the lowest level, each of the levels of language is used to convey increasingly more complex meaning and each increasing level is also more difficult to simulate in technology. This is because the higher the level, the larger the unit of analysis and the less rule-oriented the language phenomena.

6. Pragmatics:	understanding the purposeful use of language in situations, particularly those aspects which require world knowledge
5. Discourse:	interpreting structure and meaning conveyed by texts longer than a sentence
4. Semantic:	determining the possible meanings of a sentence, including disambiguation of words in context
3. Syntactic:	uncovering the grammatical structure of a sentence based on an analysis of words in a sentence
2. Lexical:	determining part of speech for each word and retrieving possible word level meanings from a lexicon
1. Morphological:	analyzing the internal components (morphemes) of words, including prefixes, suffixes, and roots

Fig.1: Levels of Language Understanding

Psycholinguistics has shown that human language understanding utilizes evidence from each of these levels of language to extract meaning when listening to or reading text. Therefore, it would make sense that language-processing systems that utilize more of these levels of language understanding will provide better capabilities.

Search Engines Today

The use of linguistics by commercial search engines today is still rudimentary. And while it is common to read in the promotional literature for a particular search engine that it utilizes NLP, when the technology is tested empirically to determine specifically what levels of language understanding are incorporated, the results overwhelmingly show that only the lowest levels of language processing (e.g. morphology and lexical) are done. The poor performance, yet overstated claims of these search engine marketers confuses both

consumers and IT Departments who are led to believe that the search engine they are selecting will include the full capabilities they attach to NLP. This understanding has resulted in much disappointment in the promise of NLP – since a search engine that uses at least one level of NLP, can claim that they are "an NLP search engine". And therein, lies the rub.

Most search engines, which state that they use NLP, simply mean that their users do not need to use Boolean logic in entering their queries. That is, they do not need to remember any specific syntax or symbols. However, this simply reduces the search engine to a simple keyword search system. Moreover, the search engines encourage their users to input one or two word queries (note the size of the box for your query!), rather than a lengthier statement which would permit the user to present their full information need. This, however, would require these search engines to have NLP capabilities beyond the word level.

The linguistic enhancements one now sees regularly in Search Engines include:

- Automatic truncation. This is mainly to enable a keyword query to match with both the plural and singular forms of a noun on the present and past tense of a verb.

- Automatic identification of proper nouns. The mechanism for doing this is simple recognition of upper case rather than anything more linguistically motivated.

- Phrase identification. Some search engines use phrase bracketers, which for the most part are based mainly on word proximity, not on recognition of meaningful noun phrases (e.g. they consider '*real nice*' as meaningful as '*real estate*').

- Concept identification. While this is a popular promotion for search engines, investigation shows that these engines rely on statistical word co-occurrence to identify concepts, rather than on true semantic understanding of concepts. While this may work at times, it is not true concept identification that can be relied on by users to recognize the true meaning of their request, no matter how it is phrased.

In general, **full-fledged** linguistic approaches have barely begun to filter into the search engine world, where the most common use of NLP is automatic truncation. While some services permit the user to enter a query without using the former idiosyncratic formats, the processing of the query appears to still be dependent on simple morphological and lexical levels of processing at most. Additionally, most systems that state they use NLP appear to perform linguistic processing on just the queries. The documents they are searching on have not usually been processed with any level of linguistic analysis.

NLP-based Information Access Technologies

There are, however, full NLP-based systems which are able to extract meaning at all the levels of language understanding for successful information access applications that are becoming available now. These are usually based on quite a few years of R & D funded by the government and provide one or more of the following capabilities: 1) a new two – staged approach to information retrieval which is more efficient and produces significantly more precise results; 2) question-answering which provides just the answer a user is looking for versus a list of documents; 3) text-mining, which applies known data-mining algorithms to free-text documents from which key elements have been automatically extracted; 4) link analysis which reveals implicit connections amongst named entities in text, and; 5) visualization technologies which enable users to quickly comprehend a summarization of multiple documents on a topic of interest. More detail

1. A 2-stage Information Retrieval system utilizes NLP capabilities in the first stage to produce a rich representation of the user's fully expressive information query, in a technology, which we refer to as Language-to-Logic (L-2-L). In this stage, the user's query which may be as lengthy and detailed as required to fully express an information need will be converted by the NLP query processor into a representation which reflects the logical relation between morphologically enriched elements in the query, and includes an identification of the key concept in the query which must be present in the documents retrieved, as well as a conceptual expansion of the concepts in the query into all their variant expressions. The L-2-L query representation is then matched against the indexes of documents produced by the various commercial web search engines with their rudimentary keyword-based indexes. The highest-ranking results (perhaps 100 to 200 documents) of this first stage of retrieval are then submitted for dynamic NLP-based processing and indexing at all the levels explained earlier in this paper. The resulting index is utilized for a second-stage matching, retrieval, and ranking of those documents, which provide **precisely** the information sought by the user.

2. Question-Answering technology is intended for those users who are in need of a focused response to their information request, and do not want to have to read and analyze full documents. NLP-based technology enables precisely this technology because it can recognize and extract relations between entities in both queries and documents. For example, a user may simply want to know when the Iran-Iraq War occurred. They do not want to have to peruse a set of full-text documents – no matter how precise the listing of relevant documents – they simply want the answer. For a more complex query such as "Who supported Iran during the Iran - Iraq War?" the NLP System's ability to extract relations (e.g. Agent, Object, Location, etc) between entities will produce a precise list of supporters. And if the question were "Who supported Iraq during the Iran-Iraq War?" a completely different list of supporters would be extracted from the documents and presented to the user.

3. Text-Mining is a specialization of Data-Mining that applies many of the data-mining algorithms – but this time to textual entities which have been automatically extracted from full-text documents and stored in a relational database (e.g. Oracle, Sybase, etc). For example, full-text documents reporting on the economy of a particular industry can be processed using NLP and the resulting textual relational database can then be processed for an uncovering of trends and patterns in the data. In fact, NLP enables Knowledge Discovery from Data (KDD) to be extended to all the information which resides in textual documents, but which can now be accessible without any manual intervention.

4. Link Analysis is an exploratory technique that has been utilized by many police investigation and intelligence agencies to manually map out the connections between individuals and organizations under investigation. NLP technology expedites the Link Analysis process by automatically extracting named entities from naturally occurring text as well as the relations amongst these entities and feeding the concept-relation pairs to a Link Analysis tool.

5. Visualization technologies need to be fed data, and NLP technologies can extract data in the form of essential elements from a set of documents for a visualization, which reveals the connections between entities – even when these occur in a number of different documents. This visualization enables users to quickly comprehend a summarization of multiple documents on a topic of interest.

When Will this Progress be Available for You?

What I have been writing about here may sound like technology that is beyond your reach. However, this is not true. All of the technologies described above have been fully implemented – and are currently in use by some sets of users. Now you may ask why aren't they available for you? Well, there are multiple reasons for this, but primary amongst these is that the true state of the art of NLP is largely unknown to those who make decisions on technology and they are wary because in the early days of Artificial Intelligence, NLP was heralded as extensively as the other technologies which comprised the new field of Artificial Intelligence. The expectations were grand and the results fell short. The vaunted goal was to start with Natural Language Processing and to advance to the point of Natural Language Understanding. And the Litmus Test of whether we had reached real Natural Language Understanding was whether the technology could do four things to a piece of text, namely:

1. Paraphrase it
2. Translate it into another language
3. Answer questions about it
4. Draw inferences from it

Now the day has arrived when NLP Systems have achieved these four goals, and we are now in the era of NLU. But you may ask 'When will I find it in the search engine I use?'

and my answer is that you will find it as soon as you express to those who are responsible for IT within your organization that you now understand what NLP can do in a search engine, and that you can recognize real NLP when you see it, and you want nothing less!

References

Feldman, S. (1996). Testing natural language; comparing DIALOG, TARGET, and DR-LINK. *Online Magazine.* Nov/Dec.

Liddy, E.D. (1998). Enhanced Text Retrieval Using Natural Language Processing. *Bulletin of the American Society for Information Science.* Vol. 24, No. 4.