Information Retrieval
via
Natural Language Processing

- or-

An Intelligent Digital Librarian

Transcription of a Presentation given by
Dr. Elizabeth D. Liddy,
School of Information Studies
Syracuse University
&
TextWise, Inc.

ASIS Mid-Year Conference
May 24, 1994

For more information contact TextWise, Inc., 2-212 Center for Science & Technology, Syracuse, NY 13244
Tel: 315-443-4456 • Fax: 315-443-5806

Information Retrieval via Natural Language Processing
An Intelligent Digital Librarian

John Garrett: This is the 1994 ASIS Midyear Meeting, *Navigating The Networks*, May 23-25, 1994, in Portland, Oregon. My name is John Garrett and I'm a member of the ASIS organizing committee. This morning's session is on enabling technologies, which is, or course, an enormously complicated and hype-ridden topic. But our speakers today will, I hope, cut through the hype and give us some substance about some of the exciting areas that are under investigation.

One of the areas that I find deeply exciting and transforming in this area is that, for the first time (at least in my memory), world class people in the computer science and electronic engineering fields and world class people in the information science and library science fields are working together to address these sorts of exceedingly difficult problems. That is a very new thing. My view is that when the history of this period is written this transformation is going to be an important step in our movement toward a really functioning and inter-operable information infrastructure.

I'd like to introduce someone you probably already know, Elizabeth Liddy, Associate Professor at the School of Information Studies, Syracuse University, and CEO of TextWise, Inc., a company she created to market a natural language text retrieval system called DR-LINK. Liz will talk today about text retrieval and natural language processing, a field that holds great promise for the future of online information gathering. Liz...

Liz Liddy: Thank you. Some of you may have been at the discussion session yesterday afternoon at which the comment was made that what is needed are "systems in which human expertise can be applied automatically to the information task." That is what I would like to talk about today -- I'm not going to talk specifically about the TextWise DR-LINK system, but rather about how information retrieval

via natural language processing (NLP), or the Intelligent Digital Librarian, offers exactly that possibility: information retrieval using human expertise.

I'm going to describe a little bit about what is needed for an information system to be truly useful (in my view, anyway), and then something about how NLP has a great potential for providing this optimal retrieval situation. I hope to show why I and many others in the field believe that NLP-based technologies offer the best approach for the future of text retrieval.

Suppose I were to stop and ask each of you to recount an experience in your past in which you were very successful in finding some information that you were much in need of. A lot of you would recall an instance in which you used an automated system -- an online catalog, or even the Internet -- and you were able to find the information quickly and easily. Others among you would probably recall an instance in which it was, in fact, an information specialist or reference librarian who was the one who saved the day for you. And then, of course, there are some of us who say "there were those instances in which it was just serendipity" -- all of a sudden you were looking for something else and there it was, exactly what you needed.

I'll get back in a minute and talk about what NLP can do for even those serendipitous situations, but if we look at those first two cases that I mentioned, I think this suggests what it is that an information retrieval system needs: an IR system must have the intelligence of a librarian together with the speed and large-scale processing capability of a computer. That is, we all need an 'electronic information assistant,' and this is what I've come to call an Intelligent Digital Librarian (IDL).

Those of you who've been in the field for a while know that IDL is not really a new name. In fact, the goal of IDL is reflected in one of the earlier systems developed by Ed Fox, Marian. But for a minute let's look at what it is that a good reference librarian does. Given that I also teach reference, this a subject near-and-dear to me.

First, a good reference librarian interprets your information need. They try and understand exactly what it is you are looking for, even when it isn't at all clear from what is said.

Second, they know the contents of their collections intimately. They know them so well they find precisely what you want, and frequently in unexpected places.

I believe that the ultimate technology for IR systems is one that can duplicate both of these human skills. I began using this phrase 'The Intelligent Digital Librarian' when I was developing an NSF proposal on the Digital Library. I made an argument there that this IDL would be able to:

- Comprehend the subtlety of a user's information need. This is *very* difficult.

- Recognize the multiple dimensions of relevance requirements presented in queries. There's much more than a main topic to a query.

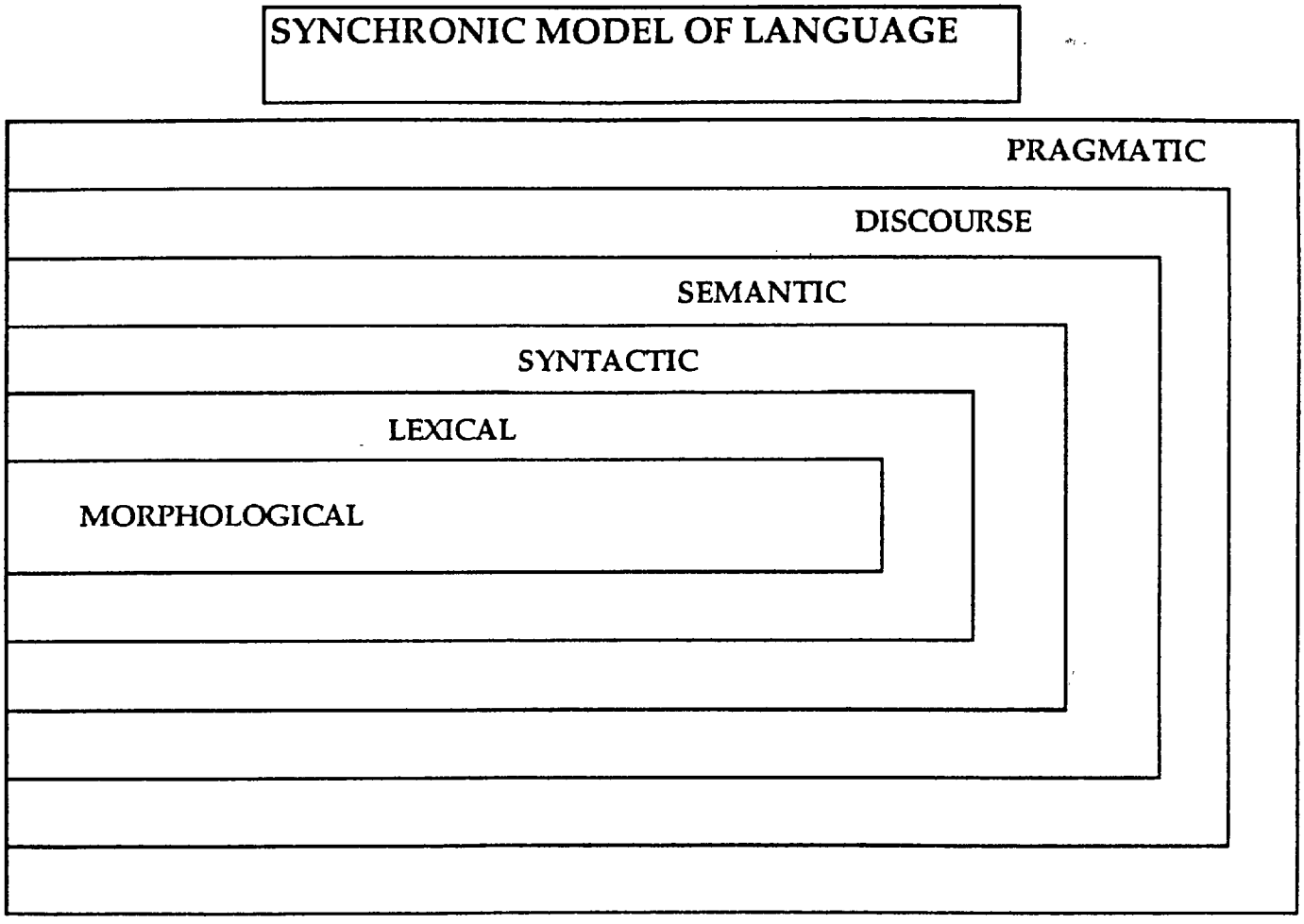
- Understand the complexity of ways in which relevant information might be expressed in various sources. There are a number of parameters that define what makes one document relevant and not another. The IDL would be able to understand the complexity of ways in which relevant information is communicated and so account for all of the variety and richness of expression that is available in a language.

- Retrieve the documents that are responsive to a query at the conceptual level.

All of these things are possible through the emerging technology of NLP, a technology whose potential is just beginning to come to realization. NLP is a technology that has the potential to provide us with "Intelligent Digital Librarians"

because it really emulates human capabilities.

At the Digital Library workshop at Rutgers last week people were saying, somewhat euphemistically, that what the desired system of the future should do is "add some intelligence" to documents. Now that's really what NLP does. It interprets the surface text strings in both queries and documents and layers on to them what we humans do: the implicit interpretations or levels of understanding. I don't want this to be too much of a tutorial, but I'm going to give you one overhead of what NLP is capable of, linguistically:



NLP is a full range of computational techniques for analyzing and representing naturally-occurring text. Not pre-constructed text. It functions at more than one level of linguistic analysis, and the purpose is to achieve human-like language processing and understanding for a wide range of particular tasks. NLP is too wide

an area for me to cover in any detail here, but I wanted to give you some sense of the range of linguistic phenomena that can be handled in an NLP system. This figure, which is sometimes referred to as the Synchronic Model of language, displays all the levels of language at which meaning is expressed. We never think about it too much, but there are many, many levels in everyday language use that convey meaning. It is only through this full range of language levels that we gain meaning:

- morphology (the simple parts of words, suffixes, prefixes, all give meaning)

- the structuring of sentences and the syntax and semantics (which helps us disambiguate the senses of a single word)

- discourse, which adds meaning for texts larger than a sentence. What is there that helps you understand the tenth or eleventh sentence in a paragraph? What is it that you carry forth from your understanding of the first sentence?

- pragmatics (how interaction with the real world impacts meaning). For example, the Cooperative Principle that underlies the pragmatics level of language is the highest level of language processing. To the question "do you know the time?" the answer "yes" in non cooperative -- we want the hour of the day.

If you look at the levels in the figure, the way they're laid out, the more exterior levels that you get to the larger the unit of analysis that you're looking at, from morphemes, to sentence, to text.

Secondly, the further out you get the less precise the phenomena you're dealing with at that level of language processing. That's because there's more free choice and variability. The farther-out levels are much less rule-oriented; they are more regularity-oriented. There are more exceptions to rules at the outer levels. The more

exterior the level you're at, the more prior levels are presumed as knowledge, or are relied upon.

In NLP as a field, and even more so in NLP as it has been used in IR, the lower levels of language have been more thoroughly investigated and incorporated in systems than the outer levels. It is these more exterior levels that make NLP still an emerging technology -- the levels of semantics, discourse and pragmatics have barely begun to be incorporated into IR systems that use NLP. The promise of NLP is that it will have knowledge at all levels of linguistic expression, allowing information seekers to express themselves naturally and with all requisite detail. The system will understand the underlying meaning of a query in all its complexity and subtlety, and understand documents at the same level. Thus both documents and queries are represented at all the levels of expression at which meaning is conveyed.

One of the most fundamental problems in information retrieval, which you might be aware of, is ambiguity. Another example from this Digital Library workshop I mentioned earlier: a group of us were watching one of the picture-image retrieval systems, and they asked for queries from the group. Someone suggested soldiers shooting a civilian. They input the query and the system retrieved several scenes of military personnel or soldiers attacking civilians, but is also retrieved one in which there was a sailor taking a picture -- shooting -- a group of rural villagers. This is very typical of how ambiguity causes problems in systems. Natural language processing promises a way of reducing ambiguity by distinguishing meaning. Another comment at the Digital Library workshop -- this came from Bob Futrelle of Northeastern -- he made the observation that for most information retrieval systems that we have, information is, in fact, encrypted in natural language. Even though human beings are able to understand, systems that do not use NLP and do not 'decode' what it is that is hidden in the language will continue to perform sub-optimally.

I'd like to look at some of the functions that we might want from a IDL, again combining many of the things we might want in a real librarian:

- Quick Reference. Users often ask for a quick fact.

- In-Depth Reference. A complex query or information need has multiple, complex parameters. For example, look at a TIPSTER query:

```
<head>    Tipster Topic Description
<num>     Number 105
<title>   Topic: "Black Monday"
<narr>    Narrative:

A relevant document will contain at least one reason why U.S.
stock markets experienced a huge price drop on 19 October 1987,
losses of equity so large that markets were said to have crashed
(the Dow, for example, lost 508 points on that day alone); the
date of the crash has become known as "Black Monday." A preferable
document would contain a detailed analysis of the crash. The best
document would link analysis of events to actions taken or
recommendations made by federal authorities or the stock markets
to prevent further crashes. NOT relevant are reports which simply
reference, without analysis, "Black Monday," such as anniversary
stories generated by the press around every October 19th.
```

This long section, under "narrative", is, in fact, the query. This is what is meant by a complex query. There are many, many parameters that would make one document relevant compared to another. This is the type of query-based system that is being developed and funded by the government. Our DR-LINK system is an example of a system that can process such complex queries.

- Understand relations between concepts. We have to think about how

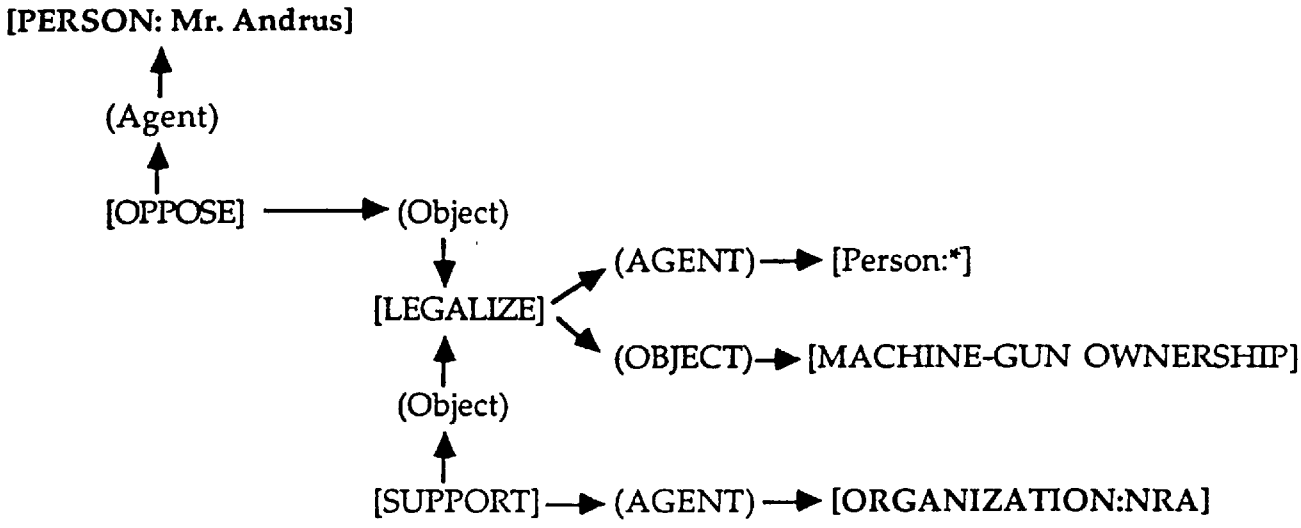
relations are recognized. Non-NLP systems usually requires a simple co-occurrence clearance, and this is not adequate. To illustrate this point take a look at this query -- again, a TIPSTER query:

```
<head>    Tipster Topic Description
<num>     Number 103
<dom>     Domain: Politics
<title>   Topic: What Backing does the National Rifle
          Association Have?
<desc>    Description:

Document must describe or identify supporters of the National
Rifle Association (NRA), or its assets.
```

Here's one of the document that was located:

```
<hl>      Idaho Feud Finds NRA Under Fire -- Peter Wiley
<so>      Wall Street Journal
<text>

One issue in the feud involves a statement Mr. Andrus made
recently opposing legalization of machine-gun ownership, which is
supported by the NRA.
```

Those of you who are familiar with how most systems operate, where simple co-occurrences are intended to indicate relations, could see why this document would be retrieved (in error). We as humans read this and understand that this is not what we're looking for. But the type of complexity that we need to understand and represent the query is something that only NLP could get for you. Consider this representation of the text:



This is a semantic representation of that text, in which the concepts are in the square brackets and the relations are in the rounded brackets. What this level of representation and meaning does for you is show the semantic relationships between concepts. In fact, Mr. Anders is not in favor of the NRA, so this document is not relevant to the query. A semantic understanding of the text is required to rule out this document for this particular query.

There are additional vital relations that make one document relevant and not another. These relations cannot be determined by co-occurrence. They are:

- Time (the temporal-ness of an event)
- Certainty
- Intentionality

- State-of-Completion

These are relations that are all recognizable and interpretable at the higher levels of NLP. Of course, we'd like an interactive refinement of the query when needed. Also the IDL should be able to explore a topic when the user can't quite define what is needed. The system should quickly separate a set of sources to browse through rather than trying to find and pinpoint the answer, to get you to the next level of understanding.

Now that we've explored what the IDL should be capable of, lets take a look at what we see in currently available information retrieval systems:

1. Queries cannot be expressed in a full, natural mode of expression. You're limited in some way either by length, type of input , or some other requirement. Systems that claim NLP may not require queries to be stated as Boolean expressions but neither do they analyze queries and documents at all the levels of language.

2. Online access does not accommodate the substantive, complex queries like the TIPSTER queries shown earlier.

3. Searching is performed at the surface with string level processing, and is based on term-level matching with no accounting for the complexity or ambiguity of language. On the network the situation is often worse: users cannot find the information they need in a natural, straightforward manner.

4. Available network tools do not facilitate precise access to information. They may be neat for browsing but complex and precise queries are not really well served. In addition, there are gem lodes of information there, but the circuitous routes which you sometimes have to follow in order to find this information takes much more

time than many of us have, and often a great deal of money.

The U.S. government recognized these deficiencies and so ARPA/NIST sponsored the TIPSTER/TREC programs. These multi-year programs have been major forces pushing our field forward. For example, the sponsors of the TIPSTER program are sophisticated users of information retrieval systems with complex, real-world information needs that are not satisfied by existing, traditional IR systems. The goal of the intelligence and military community was to push the field forward to the next level of capability in order to provide information systems for government analysts. When the programs were first announced, when the queries were first released (and those of you familiar with TREC realize that these TIPSTER & TREC queries are much the same), many people in the IR field said "this isn't what IR is about! These queries go beyond the IR queries that we're used to. These aren't the systems that we're building." Thank goodness not many people followed that response, a response that would have tied us to the past instead of propelling us forward to the future.

In fact, the TIPSTER/TREC queries do reflect the needs you see amongst many, many individuals. A list of the system requirements desired by the federal government, academic and corporate worlds are consistent. This is a long list but it is based on real requirements. These are actual tasks and jobs that need to be accomplished:

- Document Detection
- Information Extraction
- Detection and Extraction Combined
- Question-Answering
- Alerting Service
- Browsing
- Database Mining
- Composite Answers

- Automatic Summarization
- Non-Redundancy
- Easy Querying Facilities
- Automatic Semantic-Level Hyperlinks
- Zoning/Paragraph Retrieval
- Cross-Language Retrieval
- Link Analysis
- Seamless Access to Multiple Databases

A lot of you are familiar with many of these requirements -- document detection, for example.

Extraction is a relatively new function, in which there are pre-established templates developed for individual topics and the system goes in and extracts bits of information to fill these templates.

Other requirements: Question-answering (for closed-questions); alerting services (in response to standing profiles); database mining (populating a database from natural language text documents); composite answers (which is a much more complex task); combining information from multiple sources (the same way a reference librarian does for you, creating new information across multiple text sources); automatic summarization (one of the new, so-called reduction technologies that will become of increasing importance as our sources of information continue to grow); and Non-Redundancy (on the size of the databases that we are searching against it is very important that the same information not be retrieved more than once -- matching at the conceptual level NLP can help).

Easy querying facilities are important, too -- this allows users to spend less time finding information and more creative time interpreting information. Zoning, or paragraph retrieval allows short segments of interest to be retrieved from long,

diverse texts. Cross-language retrieval would allow searchers to breach language barriers to interrogate multi-lingual databases: users could input their query in one language, search across foreign-language databases, and get back inter-leaved, translated documents. And finally, seamless access to multiple databases is a high priority for many users: truly useful systems will not force the user to select which databases in which they'll search.

All of these requirements can be achieved in a system which uses real NLP.

IR system users need better search tools. If you remain unconvinced of this, let me show you one quote from a recent *NY Times Magazine,* in which an unnamed, but well-respected database service (which will remain nameless) was referred to:

New York Times Magazine, May 1, 1994.

" I've been working with XXX, a marvelous resource, although... you need to hire a special librarian to answer your question and then you never get your question answered, **all you get is the answer to the question the machine allowed the librarian to ask."**

And that's the situation we'd all like to avoid. We need to search out those information retrieval systems that comprehend text in the same way humans comprehend text -- not as keywords or phrases, but as rich discourse. Thank you.

John Garrett: Thank you very much Liz.

Liz's note on the importance of alerting services reminds me of an experience I had with the Netnews alerting service that is run out of Stamford. It works, I think, extremely well, except recently I was a little surprised because my subject query, which is 'information filtering' , turned up an anonymous article about how to

make a hashish water pipe. When I sent it back to Stamford's researchers we found there where four of five instances in the article where the word 'information' -- about how to make the pipe -- and 'filtering' --through the water -- occurred in the same sentence. So it was a perfectly valid response to my query. Unfortunately, I'm not likely to make extensive use of this subject at this point in my life, and I'm certainly going to keep it away from my kids.

THIRD INTERNATIONAL SYMPOSIUM: NATIONAL SECURITY & NATIONAL COMPETITIVENESS: OPEN SOURCE SOLUTIONS Proceedings, 1994 Volume I - Link Page

Previous	Beating the Competition: From War Room to Board Room
Next	The Digital Threat: United States National Security and Computers

Return to Electronic Index Page