Natural Language Processing for Information Retrieval and Knowledge Discovery

Elizabeth D. Liddy

Natural Language Processing (NLP) is a powerful technology for the vital tasks of information retrieval (IR) and knowledge discovery (KD) which, in turn, feed the visualization systems of the present and future and enable knowledge workers to focus more of their time on the vital tasks of analysis and prediction.

NATURAL LANGUAGE PROCESSING

First, a definition of NLP. Natural language processing is a set of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.

The goal of researchers and developers of NLP is to produce systems that process text of any type, the same way which we, as humans, do—systems that take written or spoken text and extract what is meant at different levels at which meaning is conveyed in language. NLP is used for a wide range of tasks or applications. This discussion will focus on two particular tasks, namely information retrieval (IR) and knowledge discovery (KD).

Figure 1 shows the levels of language processing at which cognitive linguists hypothesize that humans understand or extract meaning. An interesting point to note is that, while meaning is frequently thought to be conveyed at the level of language represented as "semantics," the following explanation will clarify how meaning is in fact conveyed and how we, as humans, extract meaning at every level of language, not just at the semantic level.

Synchronic Model of Language

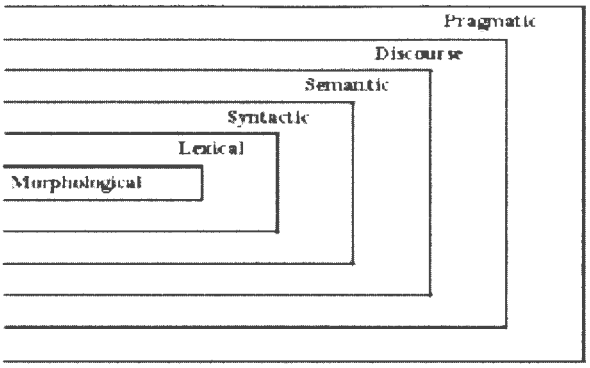

Figure 1. Levels of Language at which Meaning is Conveyed

- The *morphological level* has to do with the smallest units of meaning in language, namely morphemes—the smallest meaningful pieces of words. For example, the morpheme "ed" at the end of a verb tells you that the action took place in the past, not that it will take place in the future. Additionally, simple things like adding the morpheme "un" to "lawfully" drastically changes the meaning of the word.
- The *lexical level* is concerned with linguistic processing at the word level and includes such processing as part-of-speech tagging. When humans hear or read a sentence, they determine whether a word that can function both as a verb and as a noun is either a verb or a noun in that particular sentence and knowing that helps them disambiguate the meaning of the word.
- The *syntactic level* is where order and the arrangement of words within a sentence conveys meaning. For example, the sentence "Clinton beat Dole" contains the same words as "Dole beat Clinton," but the simple ordering of those words conveys a world of difference in meaning.
- The *semantic level* is concerned with understanding the meaning of words within context—i.e., humans are able to unambiguously understand words when they hear them or read them in a sentence even though many words have multiple meanings. For example, in the English language, the most commonly occurring verbs each have eleven meanings (or senses) and the most frequently used nouns have nine senses, but humans can correctly select the one sense or meaning that is intended by the author or speaker.
- The *discourse level* is concerned with units of text larger than a sentence. Discourse is a newer area of linguistic applications, having begun as an area of linguistic study in the 1970s. Discourse linguistics is concerned with the linguistic features that enable humans, for example, to under-

stand the eighth sentence in a paragraph partly because of the meaning they extracted from the first to seventh sentences. Discourse is also concerned with utilizing the fact that texts of a particular type (also known as a genre) have a predicable informational structure and that humans use this structure to infer meaning that is not explicitly conveyed at any of the other levels in the model.

- The *pragmatic level* is concerned with the knowledge and meaning that we assign to text because of our world knowledge. For example, the phrase "Third World Countries" does not just mean those three words to a reader. Pragmatic knowledge brings in a lot of other understanding, such as which are the Third World Countries and the general socio-economic conditions in these countries.

A further fact of interest is that the more exterior the level of processing (as shown in Figure 1), the larger the size of the unit being analyzed, ranging from a part of a word, to a word, to a sentence, to a paragraph, to full text. And as the size of the unit being analyzed increases, processing rules get less precise—i.e., there are fewer rules to rely on, just regularities. For example, there are precise rules about how to write a grammatically correct sentence, however there are only regularities that explain how a newspaper article is written. And so natural language processing is more difficult to do at the more exterior levels as it is not simple rule writing as one would be able to do at the lower levels. This fact explains why many systems limit their language processing to the lower levels and most of them do not, in fact, include the higher levels—i.e., real semantic, discourse, and pragmatic processing. In conclusion, a full NLP system extracts meaning from text at all the levels of language at which humans extract meaning.

INFORMATION RETRIEVAL

The goal of information retrieval is to provide a user with documents that fulfill the user's information need. This involves system capabilities for indexing (how to represent the content of documents, including appropriate weighting schemes); representing users' queries (including both the means provided to users to express their queries and the complexity of the internal representation of the query); matching algorithms to optimize true similarity between a query and relevant documents; techniques for effectively presenting retrieved results, including summarization across documents and visualization of results; and techniques for improving query results based on users' relevance assessments via relevance feedback.

Current IR approaches can be classified as statistical (vector, probabilistic, inference, neural net), linguistic (ranging from very simplistic word stemming to complex semantic processing), or a combination of both

statistical and linguistic. Further explanation of how NLP is used for IR is presented in a later section on applications.

KNOWLEDGE DISCOVERY

Knowledge discovery is the area of research and development involved with the computational process of extracting useful information from massive volumes of digital data. The goal of KD is to map large quantities of low-level data into a more abstract form so that the patterns in the data can be explored and inferences drawn from them. KD provides a range of techniques and methodologies to extract knowledge automatically from these sources.

NLP offers the field of knowledge discovery the ability to go beyond the limited information that is stored in structured or relational databases, which has been the source of data for knowledge discovery to date. However, most of the world's knowledge resides in free text form—that is, unstructured, naturally-occurring text such as encyclopedias, newspapers, textbooks, and so on—so NLP provides the means to process, annotate, and extract information from text to produce new resources for knowledge discovery.

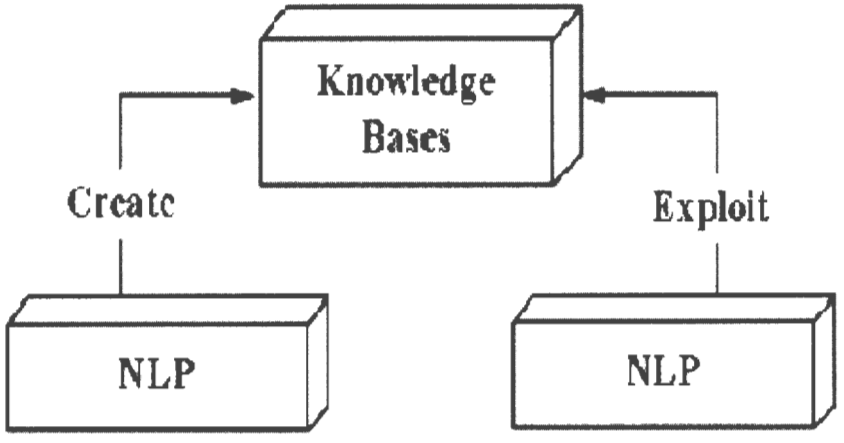

Figure 2. Uses of NLP in Knowledge Discovery

In fact, as seen in Figure 2, NLP is useful in two ways for KD. First, as stated earlier, it can be used to automatically create knowledge bases from free-form text using the extracts/annotations that NLP does on the text. The knowledge bases constructed in this manner can then be manipulated by a wide range of different statistical data mining systems. Or, second, NLP can itself be used to exploit those knowledge bases. That is, NLP enables users to explore the knowledge stored in these databases simply by asking very straightforward queries. A Knowledge Discovery

System that uses NLP to extract full meaning from a query can then do knowledge discovery at all the linguistic levels represented in the stored knowledge base.

NLP FOR KNOWLEDGE PRODUCTS

The full processing model that forms the basis of the various technologies described below consists of four steps (see Figure 3). That is, the system starts with raw text, performs NLP on it, and produces extractions that are then converted into a semantic representation for use in a knowledge product.

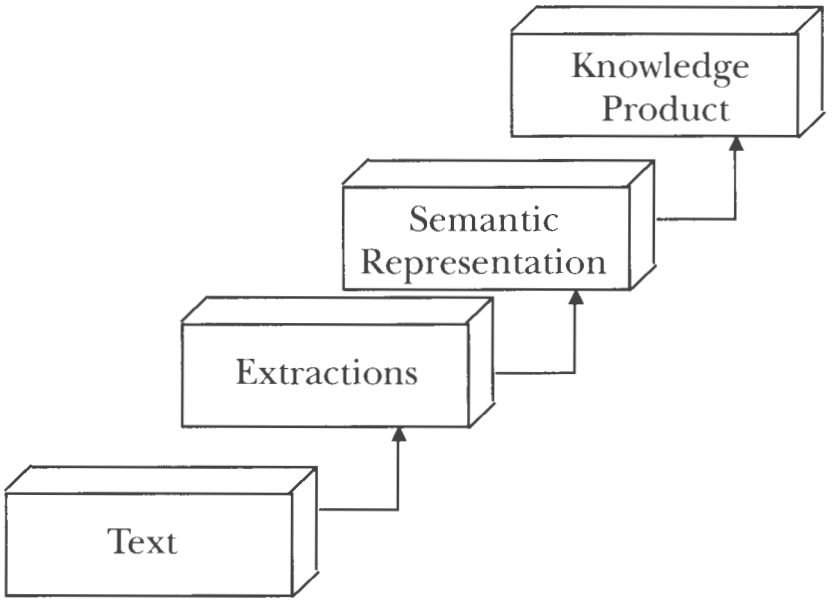

Figure 3. Steps in NLP for Knowledge Products

Step one assumes naturally occurring free-flowing text such as newspaper or journal articles, books, reports, TV transcripts, and so on. The second step in the ladder is extraction. For instance, our DR-LINK Information Retrieval System does part-of-speech tagging, phrase and clause bracketing, standardizing and categorizing of proper nouns into one of fifty-seven categories, and indicates parameters such as present, past, or future and distinguishes between fact and opinion.

The third step is to map these extractions into a semantic representation. The semantic representation used in an application may be either a semantic net, frames, logic, conceptual graphs, or whatever best suits the requirements of the application. In our CHESS and KNOW-IT Systems, we use concept-relation-concept triples. For example, in

• 03/01/95
Robert Dole → (AFFL) → Senate
→ (AGE) → 71
→ (ORGN) → Russell Kansas
→ (IS_A) → clear Republican Party front–runner

. . .
• 03/21/95
Robert Dole → (IS_A) → presidential candidate
→ (AGNT) → seek
→ (OBJ) → repeal of assault-weapon ban

. . .
• 03/23/95
Robert Dole → (IS_A) → rival
rival → (OBJ) → Phil Gramm
Phil Gramm → (IS_A) → Senator
→ (ORGN) → Texas

Figure 4. Concept-Relation-Concept Triples from Newspaper Text

CHESS, we build a chronological record of people, companies, and organizations from daily newspaper reports. Figure 4 shows the concept-relation-concept triples from three newspaper excerpts about Robert Dole.

The fourth level is the actual knowledge product that results from the three prior steps. So in our CHESS System, having extracted these concept-relation-concept triples regarding Robert Dole, CHESS then produces a semantic network for use in a product for browsing, exploration, and knowledge discovery.

NLP-BASED APPLICATIONS

The section above provides a general model of what is done in various NLP-based information systems. Some processes are common to all the products while some products require a specific level of processing or a particular representation. In this section, the focus will be on how NLP is used in a few specific applications.

DR-LINK

DR-LINK (Document Retrieval using LINguistic Knowledge) is a powerful full-NLP information retrieval system that encompasses all levels of language processing described earlier. We developed DR-LINK initially under the auspices of DARPA's TIPSTER Program followed by commercial development by TextWise, LLC, and Manning and Napier Information Services. DR-LINK is now commercially available and is used by government agencies, businesses, law firms, and other professional organizations.

In DR-LINK, documents and queries are processed through a series of modules in the system, each of which add another level of representation of the content by extracting and annotating meaning at the various levels of language processing explained in the earlier sections. Briefly, the raw documents are part-of-speech tagged; the Subject Field Coder (which is explained in more detail in a later section) produces a disambiguated semantic vector representation of the subject content of each document and query; the Text Structurer uses discourse linguistics to understand whether an event will occur in the future versus the past and can also distinguish fact from opinion; the Proper Noun Interpreter recognizes and categorizes all proper nouns into one of fifty-seven categories; and the Phraser is the module that creates the list of synonymous terms for the terms/phrases in the text.

Finally, an integrated matcher takes the ranking of each document as suggested by each module in the system and combines this evidence to produce a relevance ranked list of documents for the query.

Query—137

Document will report on the proposed building of a new or the expansion of an existing theme park by a U.S. corporation in the United States or overseas.

DR-LINK Retrieved Document:

Wall Street Journal

10/01/87

Six Flags Corp announced plans to manage and operate an amusement park on Spain's Costa del Sol.

Scheduled to open in 1990, the park is intended to form part of a 107,000 acre tourism complex costing about $575 million.

The company, a unit of closely held Wesray Capital Corp. of Morristown, NJ, said Spain is anticipating an increase in tourism because of the 1992 Barcelona Olympics and Seville World's Fair.

Figure 5. Sample Query and Relevant Document Retrieved by DR-LINK

Figure 5 shows a query and a document that DR-LINK found relevant. Note that there is no term match between the query and this document, and it is because of these multiple levels of processing of the document and of the query that DR-LINK is able to do such a human-like NL understanding match.

INFORMATION EXTRACTION

Information extraction, based on NLP, adds to the capabilities of IR applications. CHESS (CHronological information Extraction SyStem) is a

system we have developed under Phase I and Phase II SBIRs from Rome Lab. CHESS extracts concepts and relations from newsfeeds, such as API, and automatically constructs complex in-depth historiographies that track people, companies, organizations, or other entities over time. The system does full NLP to extract concept-relation-concept triples and construct conceptual graphs. CHESS does information extraction (IE), but it also goes beyond IE. Given the fine level of processing and representation, CHESS can do something that is a bit beyond what systems that rely on co-occurrence, or Boolean representation, or any type of statistical processing can do.

Tipster Query—93	Wall Street Journal 12/21/87
What backing does the National Rifle Association have? Document must describe or identify supporters of the National Rifle Association (NRA), or its assets.	Idaho Feud Finds NRA Under Fire—By Peter Wiley . . . One issue in the feud involves a statement Mr. Andrus made recently opposing legalization of machine-gun ownership, which is supported by the NRA.
Figure 6a	**Figure 6b**

Figure 6c

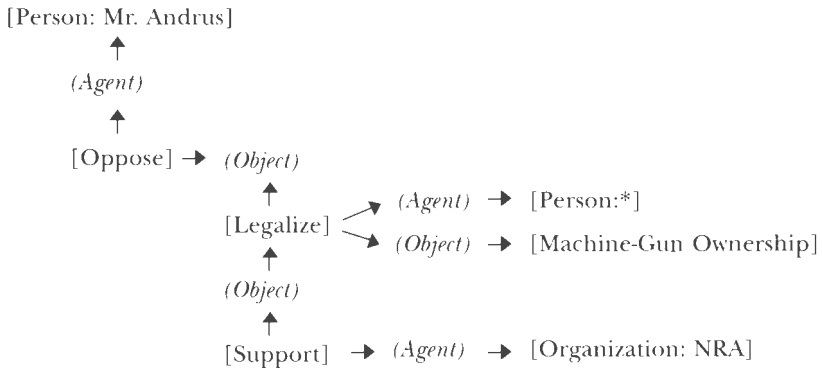

Figure 6. Query, Document, and Conceptual Representation in CHESS

For example, given the TIPSTER query in Figure 6a, most humans would judge the document in Figure 6b as not relevant. But most systems, either because of co-occurrence or proximity measures or the type of matching that they do, would find this document to be relevant. However, given the fuller conceptual semantic representation produced by CHESS in Figure 6c, CHESS understands that Mr. Andrus is an agent who opposes the legalization of firearms which is supported by the NRA. So CHESS is able to tease that out and say: "No, this is not a relevant document" much as human judges can.

Figure 7 provides examples of the types of questions which CHESS, a

system that does NLP-based IE, could answer from knowledge that the system has automatically extracted from daily newsfeeds. Specifically, CHESS can do historical or chronological tracking of people over time—recognizing and matching on relationships between two entities, such as the relationship between H. Ross Perot and the Democratic Party. But perhaps an even more powerful capability is that a user can start with queries that are about relationships in which the entities are not known, but the user wants to know which entities exist in particular relationships with each other. This is what is done in scenario analysis—it is much like defining a frame and asking what are the entities that have been observed in text to have the necessary characteristics to fill these slots in the frame.

And in an emerging scenario, CHESS enables the user to make use of knowledge of certain events the system has tracked over time. The system can learn the set of events that lead to a particular outcome and then the user can ask whether these other entities are following that same track. This capability is very useful for government, for business, for competitive intelligence, or for anyone who is tracking anybody or anything.

- Historical/Chronological
  —*When did H. Ross Perot resign from GM?*
- Association Questions
  —*What is the relationship between H. Ross Perot and the Democratic Party?*
- Scenario Analysis
  —*Which political figure with military connections is operating in an unstable, third world country?*
- Emerging Scenario Analysis
  —*What company is going to file for Chapter 11?*

Figure 7. Questions Which CHESS Will Aid in Answering

Knowledge Discovery

The next example of a knowledge product is the Know-It System, which broadens the CHESS Information Extraction work to knowledge discovery by extending the use to which the automatically extracted concepts and relations are put. Know-It will use full NLP capabilities to build ontologies that represent the complex knowledge contained in source texts. With ontologies, because of their subsumption relations, one can make very powerful inferences from the knowledge stored in the ontology.

The system will be provided with field manuals and, by doing natural language processing of them and using Know-It to automatically construct ontologies of different situations, the system will be able to compare dynamically a field report as it comes in on a certain situation which is represented as an event-based ontology in the pre-constructed

reference ontology to say where it differs. So a user would be able to do retrospective analysis, or lessons learned, to review what the difference is in the ontologies that represent situations that were successful as compared to those that were unsuccessful.

Furthermore, Know-It will provide decision-making information support by producing to-the-point answers to users' questions and by permitting them to test their hypotheses. Additionally, Know-It will enable collaboration by providing a networked tool with multiple views of the knowledge space at different levels of specificity and from multiple viewpoints.

Information Abstraction

The next knowledge product is information abstraction. For this we will look in more detail at one of the modules of DR-LINK, namely the Subject Field Code (SFC) module, which produces a weighted semantic vector representation of each document and query. The SFC representation fits somewhere between controlled vocabulary representation/searching and free-text representation/searching. The goal of the SFC vectors is to produce a summary level representation of the semantic content of a document, but not at either the automatically indexed individual term level or the manually assigned, controlled vocabulary level. The SFC vectors provide an abstract level of representation by using a set of about 700 subject codes and representing each document as a weighted vector across all of these codes. The most powerful aspect of these codes is that they handle both synonymy and polysemy, the two most difficult problems in NLP—synonymy—the fact that one concept can be represented by many terms—and polysemy—the fact that one term can represent many concepts. The SFC module is able to handle these complexities because it disambiguates using the three sources of evidence that psycholinguists say that humans use when disambiguating. Furthermore, the SFC does not require human analysis or large training corpora. The system processes text across many domains, and it does it very quickly and automatically.

Most NLP systems make use of a lexicon in which various types of information are stored for each word or each sense of a word, such as part-of-speech, a definition, how to form the plural, and so on. DR-LINK has a lexicon that contains the SFC for each sense of a word. So, for instance, if you looked up a common word such as "instrument," the system would consult the lexicon and report that "instrument" has a medical sense, a hardware sense, a dental sense, a musical sense, and a general sense. The system would then disambiguate among the senses the same way that a human would. The system is able to disambiguate at almost a 90 percent accuracy level for determining in a document what the intended sense (or SFC) was. So having done that, the system is able to create a weighted vector by simply normalizing frequencies across the document.

CONCLUSION

In conclusion, I will repeat something I heard said by Bob Futrelle from Northeastern University at one of the Digital Library Conferences and which makes good sense to me and that is: "For most IR systems, information is encrypted in natural language and NLP is the code breaker."